调查方法

罗伯特 M. 格罗夫斯 (Robert M. Groves)
弗洛伊德 J. 福勒 (Floyd J. Fowler Jr.)
米克 P. 库珀(Mick P. Couper)
詹姆斯 M. 勒普考斯基 (James M. Lepkowski)
艾利诺 · 辛格 (Eleanor Singer)
罗杰 · 图兰吉 (Roger Tourangeau) 著

邱泽奇 译







WILEY

Survey Methodology, Second Edition. by: Rorbert M. Groves, Floyd J. Flowler Jr, Mick P. Couper, James M. Lepkowsky, Elean Singer, Roger Tourangeau, ISBN: 978-0-470-46546-2

版贸核渝字（2017）第001号。

图书在版编目（CIP）数据

调查方法 / （美）罗伯特M.格罗夫斯（Robert M. Groves），等著；邱泽奇译. —重庆：重庆大学出版社，2016.11

（万卷方法）

书名原文：Survey Methodology

ISBN 978-7-5689-0206-9

Ⅰ．①调…　Ⅱ．①罗…②邱…　Ⅲ．①调查方法－研究　Ⅳ．①C31

中国版本图书馆CIP数据核字（2016）第254159号

调查方法

DIAOCHA FANGFA

罗伯特M.格罗夫斯（Robert M. Groves）

弗洛伊德J.福勒（Floyd J. Fowler Jr.）

米克P.库珀（Mick P. Couper）

詹姆斯M.勒普考斯基（James M. Lepkowski）

艾利诺·辛格（Eleanor Singer）

罗杰·图兰吉（Roger Tourangeau）　著

邱泽奇　译

策划编辑：雷少波

责任编辑：文　鹏

版式设计：雷少波

责任校对：秦巴达

责任印制：赵　晟

重庆大学出版社出版发行

出版人：易树平

社址：重庆市沙坪坝区大学城西路21号

邮编：401331

电话：（023）88617190　88617185（中小学）

传真：（023）88617186　88617166

网址：http://www.cqup.com.cn

邮箱：fxk@cqup.com.cn（营销中心）

全国新华书店经销

自贡兴华印务有限公司印刷

开本：787mm×1092mm　1/16

印张：23

字数：477千

2017年2月第1版　2017年2月第1次印刷

ISBN 978-7-5689-0206-9

译者前言

《调查方法》一书的翻译始于10年前。

2006年暑假期间，北京大学和密西根大学联合在北京大学举办调查方法培训班。学员主体是部分高校承担社会调查课程教学的教师，另有部分在校学生。主讲教师是密西根大学社会研究院（Institute for Social Research）的教授James M. Lepkowski——《调查方法》的作者之一，使用的教材则是《调查方法》的第一版。在教学中，我的角色名义上是中方合作教师，实质上是助教。

在学术路径上，我最初的学习起始于与《调查方法》作者们尊敬的老师Charles F. Cannell和Rensis Likert相同的行当，也实践过农业科学的田间调查方法；阴差阳错，1986年，我开始在讲台上讲授社会调查方法；1994年还接受过Leslie Kish的短期教诲。在经历过这些之后，我自以为对调查方法不仅有理论积累，也有经验积累，还设计和实施过大规模的社会调查。

可是，当我安静地在教室里听James讲授《调查方法》、阅读他留下的文献后才渐渐发现，我正是他们指称的只讲技术不讲原理、照猫画虎的教师中的一员。的确，多少年来，我们过分地专注于社会调查技术层面、计算层面和方法的“形似”，忽略了技术和计算背后的理论逻辑，尤其忽视了调查方法的精髓：如何在预算约束条件下综合运用调查方法和管理手段，最大限度地减少调查误差，提高调查数据的质量。

正是这样的触动，使我有了翻译的想法并立即付诸行动，组织了几位有热情的年轻人一起努力。或许是暑期班的时间太短，或许是我的组织能力有限，第一版的翻译流产了。一份残缺的译稿在随后一系列事件的冲击中被我放下了。

2006年秋天，我接受北京大学的委托，创办北京大学中国社会科学调查中心（简称“调查中心”），并创立中国家庭跟踪调查项目（Chinese Family Panel Studies，CFPS）。在这一过程中，重庆大学出版社的雷少波编辑到北京大学来组稿，我强烈推荐了《调查方法》。理由是，在面世的调查方法教材（抽样和问卷）中，它是对调查方法的原理和操作说明呈现最清晰的一部教材。雷先生慧眼，马上接受了我的推荐，同意把《调查方法》列入他组织的“万卷方法”丛书中，并期待着我能尽快交稿。

我原以为自己能挤出时间和精力来完成译稿，但事实是，我低估了创办一个机构和初创一个项目的工作量和困难。从2006年创办机构开始到2010年CFPS初访调查结束的整整5年时间里，尽管我每天工作至少十五六个小时，可除了调查中心的组织工作和CFPS的设计、实施、管理等工作以外，我再没有时间和精力写作和翻译。这5年，是我学术写作的荒芜期，《调查方法》的翻译自然也被耽搁了。

到2011年把《调查方法》的翻译捡起来的时候，原书已经出版第二版了。非常感谢雷先生的耐心，此时，他还在等我的译稿。出于知识更新的考虑，我们商定放弃第一版的翻译，直接翻译第二版。这就是《调查方法》译本的由来。

对于译作的品味，严复先生曾经提出过一个标准：信、达、雅。在三十年的翻译实践中，我的体会是，对技术文献的翻译即使做到信

也不易。理由是，它不仅需要译者能透彻地理解技术的原理和操作，或具有和作者同样的技术背景和能力，还需要能用中文准确地再现出来。对于多人参与的著作而言，难处尤其凸显，这是因为一位译者不可能对每一位作者的学术背景、能力、风格都有准确的把握。这一点，读者们从本书译文中将会有清晰的感受。坦率地说，《调查方法》的作者们都不善于通俗表达，原著中的语言与表达更多地侧重了技术准确性，忽略了阅读和理解的通俗性。如此，翻译要做到信，更加困难。

如果说在翻译第一版时我还不是那么有把握的话，那么，在翻译第二版时自认为对书的精髓有了相对准确的了解。了解一方面是来自于在翻译耽搁期间我创办了调查中心和创立了CFPS，再次实践和体验了调查各个环节涉及的方法和操作。更重要的是，我与作者中的几位有了比较深入的交流，尤其是第一位作者——Robert M.Groves，他是密西根大学社会研究院的教授，也是美国2010年人口普查的负责人。在他的办公室，我们曾经就应答率对数据质量的影响进行过深入交流。即使如此，我依然不能保证我的翻译完全达到了信的标准。

需要说明的是，读者拿到的译本在版面上与原著有一些差异。譬如原著每一章的核心概念以页边文字出现，在中文版中改为了在正文中以不同字体出现；原著书尾的索引，由于中文版面索引排版困难而暂时删除了，等等。不过，原著的精髓在译本中则完整地保留了。

对《调查方法》内容的组织和阅读建议，作者们在第一版序言中有清晰的说明；对第一版与第二版的区别，在第二版序言中也做了交代，这里不再赘述。

在第一版翻译中，傅强、穆峥、王志理、张磊、邹智敏、左冬梅等各自提交了一章的译稿。在第二版翻译中，为使译文风格一致，我放弃了第一版已有的初译稿。在第二版初译稿完成之后，北京大学社会学系的部分研究生进行了校读；在清样出来之后，北京大学2016年秋季学期修读“社会调查与研究方法”课程的学生再次进行了校读。即使如此，译本中依然可能存在差错。读者在阅读中如发现任何差错，文责在我。

感谢所有为《调查方法》译本面世作出了贡献的人们，尤其感谢重庆大学出版社和本书的策划编辑雷少波先生。

2016年12月于北京皂君庙

第2版序言

我们非常高兴大家接受了《调查方法》的第1版。这本书已经在世界各地被当作教材，也被翻译成了多种语言的版本。有的老师和学生还殷切地指出了一些章节的弱点甚至错误，有的还对教材的改进提供了很好的建议。

此外，作为调查方法专家，我们都积极参与了实地调查研究，我们也因此越来越多地知道教材中有些知识已经过时。在搜集数据的新方式不断产生的情况下，关于无应答与调查质量评估的章节尤其如此。

基于这些原因，作者们聚集在一起，同意更新部分章节的内容，以反映近期的发展。正如读者会看到的，第3章增加了用于移动电话和互联网调查的抽样框讨论。第4章讨论抽样设计时整合了例子用于讨论家庭户内个体的抽样。第5章更新了移动电话和互联网调查的内容。第6章讨论调查无应答，变动是最大的，反映了一些新的洞见，如无应答率与无应答误差之间的关系。第8章针对评估调查质量问题，强调了问卷设计研究中的一些新发现。第11章讨论调查中的伦理问题，特别强调了隐私、知情同意、保密议题。其他的章节尽可能地反映了近些年方法研究领域的进展，特别是涉及调查误差时。所有章节中，50%的习题都依据教师使用教材中的反馈进行了调整。

两个助理协助整理了这一版的手稿，Michael Guterbock和Kelly Smid。密西根大学的一些博士研究生（Ashley Bowers，Matthew

Jans，Courtney Kennedy，Joe Sakshaug，以及Brady West）阅读了初稿。在我们与Wiley签订合约时，我们仍然邀请了Lisa Van Horn作为我们的出版编辑。是上述所有人的努力，使得这一版能够成功出版。谨此致谢！

和上一版一样，我们将用本书的销售所得来支持刚刚踏入调查方法领域的人，稿酬将直接捐给Rensis Likert调查方法研究基金会，并直接用于调查方法的研究生培养。

Ann Arbor，Michigan	ROBERT M. GROVES
Boston, Massachusetts	FLOYD J. FOWLER, JR.
Ann Arbor，Michigan	MICK P. COUPER
Ann Arbor，Michigan	JAMES M. LEPKOWSKI
Ann Arbor，Michigan	ELEANOR SINGER
College Park, Maryland	ROGER TOURANGEAU

2009年3月

第1版序言

我们写这本书有一个特殊的目的。我们都是调查方法的专家，但也都是学生——在被称为“调查研究”的诸领域中数据搜集和分析的实践方面。今天人们所看到的调查，大约有60～80年的发展史。过去20年发展出来的一系列理论和原理，为设计、执行、评估调查活动提供了统一的标准。这就是大多数时候所说的“总体调查误差”范式。这个框架指导了调查质量的研究，形塑了专业化的调查方式。在调查研究领域兴起的这些研究，就是这里所说的“调查方法”。

当然，我们也逐步注意到教科书上的“调查”和作为科学发展的“调查”之间的差别。许多教科书所讲的调查研究，只是关注工具的应用，而忽视了工具背后蕴含的理论与科学。一些教科书教学生如何做事，但却不提供方法研究领域的背景支持。简而言之，有些书强调的是“怎么做”调查，但却忽视了“做”背后的科学。

我们认为最有害的是，人们读方法书的目的是为了完成一项任务，认为按照书上的指引去做，就能保证获得高质量的调查结果。而我们的看法恰恰相反，我们认为调查是针对特定目的和特定目标群体，对原则的适用性应用；就特定的目的和目标群体而言，具有唯一性。

当我们为“调查方法联合项目”（JPSM）准备一个学期的研究生课程时，这些问题变得特别重要。调查方法联合项目是美国联邦统计局资助的一个研究生教育项目。参加这个项目的学生大多在其他领域

（如经济学、统计学、心理学）接受过高层次的教育，但却没有机会从事过实地调查。为此，从1998年秋季开始，我们计划了一个14周的教学课程，包括作业和考试，但却为没有教材和作业而苦恼。

为此，我们筹划一本教材来阐述调查设计的基本原则，反映过往调查方法研究领域的成果以及与高质量调查相关的决策指导。我们想纳入习题，以帮助学生理解这个领域的知识。我们希望传递的信息是，这个领域是以实验和其他研究发现为基础的领域，调查设计不仅仅是一系列研究想法和操作权衡，更是研究积累的后果。

草拟这本书花费了我们几年的时间。在写了最初的几章以后，我们感到灯枯油尽，直到我们的同事Nancy Mathiowetz唤醒我们，给我们以力量回到了正常的轨道。为此，我们非常感激她的鼓励。

文稿中有相当一部分内容得益于我们的学生兼同事的批评。2003年，Maria Krysan和Sue Ellen Hansen在密西根大学调查研究中心（SRC）的暑期调查研究技术班上，讲授了“调查研究技术导论”，才使这本教材最终定稿。非常感谢两位教员帮助改进了教材。从两位教员和班上学生的批评中，我们学到了很多。这些学生是：Nike Adebiyi，Jennifer Bowers，Scott Compton，Sanjay Kumar，Dumile Mkhwanazi，Hanne Muller，Vuyelwa Nkambule，Laurel Park，Aaron Russell，Daniel Spiess，Kathleen Stack，Kimiko Tanaka，Dang Viet Phuong，and Christopher Webb。

平心而论，这本教材也深切地反映了那些授课教师们对我们的教育。需要特别说明的是，作者们相互都是朋友，也都是Charlie Cannell的学生，有的是正式的，有的是非正式的。Charlie F.Cannell在美国农业部项目调查司师从Rensis Likert开始调查生

涯。1946年，在建立密西根大学调查研究中心的时候，Cannell又与Likert和其他同事相遇。他是调查研究中心第一任实地执行主管，在调查方法领域具有长期的经验和卓越的贡献。为了纪念Charlie和他的工作，社会研究院（ISR，SRC是ISR的一部分）在调查方法部分专门设立了Charlie F. Cannell基金。这本教材销售的所有收入将会全部捐给基金，用于支持年轻学者们在调查方法领域的研究工作。我们想，应该没有比这更好的选择了。

这本教材适用于已经修读过1～2门统计学的人，基本技能包括读懂统计符号，如求和的符号、期望值符号，以及简单的数据演算。有些章节含有回归、logistic回归模型的量化分析，没有线性建模知识的人在这些章节需要一些其他课程知识的帮助。

全书共有12章，按照在JPSM“调查方法基础”一学期课程的教学顺序排列，包括可能从课后阅读文献中挑选出来额外布置的作业。

第1—2章（“调查方法导论”和“调查中的推论与误差”）是概念性的。第1章列举了6个调查实例，这6个例子将会贯穿全书，用于说明各种原理和实践。老师可以找到这些调查的网站来补充教材介绍的不足，并用于一些关键设计特征和调查产出的课堂讨论。

第2章呈现了涉及调查误差的关键因素。当然，课程开始的时候，我们发现，用例子更能让学生理解涉及调查误差的关键因素。正如我们看到的，各种调查的一个显著特征就是通过调查来对总体进行统计描述。尽管可以使用计算机进行统计计算，但调查专家须理解统计背后的含义是十分关键的。因此，本书在进行统计量计算的同时进行概念的讨论。

可以将第2章当作统计概念课程，一旦掌握了基本统计概念，学生们在后面的学习中就会更加得心应手。

从第3章（目标总体、抽样框，以及覆盖性误差）开始，每一章将讨论调查误差中的一种误差以及现有方法研究给予的解决原则。这些章节的讨论都是在最佳研究实践的基础上展开的。我们常常发现，刚开始学习调查方法的学生总是认为他们的某个专门设计是最好的。第3—11章展示的资料试图说明，所谓最佳实践，是需要使用调查方法进行科学研究才能得出结论的；“认为”的价值是很有限的，除非有研究结论作支撑。在这类研究中，有些也没有特别的发现。因此，从事调查的学生一定要阅读过去的方法文献，并不断地作变异性研究，如此才能确定是否是好的设计。本书有两个工具可帮助学生理解这样的观点：一是相关论题下的参考文献；另一个是每一章文本框例举的经典研究，说明了设计、发现、局限以及影响。从后面列举的“进一步阅读资料”还可以找到这些研究的全文，并可以用于课题讨论。

第4章（抽样设计与抽样误差）比其他章节运用了更多的统计概念。当课堂上有许多人需要阅读和理解统计概念时，我们会建议他们去读Kalton的一本小册子《调查抽样导论》（Sage，1983）。有些情况下，我们甚至会花3周的时间来讲覆盖性和抽样的章节。

第5—10章，每一章是一周的内容。我们发现，非常重要的是要强调覆盖性误差与无应答误差具有同等重要性。我们还强调了把阶层间关联原则运用到整群样本效应和访员方差。

第11章（涉及科学性的原理和实践）不仅包括了敏感性训练，也包括了涉及研究人类主题的伦理问题的概念框架和近期关于调查数据

隐私性分析的理论与实践。当然，我们说明了研究和判断如何能影响伦理相关的决策。

第12章（调查方法常见问题）的风格完全不同于其他章节。这是一个传统，在考试之前，总是要花时间来解答问题。我们发现，学生们这时提出的问题，总是试图从更广的视角整合课堂学到的知识。因此，我们写了“常见问题”，包括一般性的问题及其答案。

编辑Sarah Dipko和Sonja Ziniel的智慧使本书得到了极大的改进。Adam Kelley帮助用计算机制作了图表。Wiley出版社的出版编辑Lisa Van Horn具有极好的节奏感。在此，对他们表示衷心的感谢。

写这本书很有乐趣，一方面表达了我们对一些关键研究领域的观点，同时也争论着如何形成调查方法的知识领域。我们希望你们也会有同样的乐趣。

Ann Arbor, Michigan	ROBERT M. GROVES
Boston, Massachusetts	FLOYD J. FOWLER, JR.
Ann Arbor, Michigan	MICK P. COUPER
Ann Arbor, Michigan	JAMES M. LEPKOWSKI
Ann Arbor, Michigan	ELEANOR SINGER
College Park, Maryland	ROGER TOURANGEAU

2004年3月

致谢

作者衷心感谢原图表版权所有者允许《调查方法》引用其图表。《调查方法》引用的图表及版权如下：

1. 图1.5c ，来自Mokdad，Ford，Bowman，Dietz，Vinicor，Bales，and Marks（2003），美国医学会授权．Copyright © 2003.

2. 表5.1 ，来自Groves（1989），John Wiley and Sons授权．Copyright © 1989.

3. 图6.5 ，来自Groves（2006），周边文字来自Groves and Peytcheva（2008），美国公共舆论研究学会授权．Copyright © 2006，2008.

4. 此处文本框中表 ，来自Merkle and Edelman in Groves，Dillman，Eltinge，and Little (2002)，John Wiley and Sons授权．Copyright © 2002.

5. 图7.2 ，来自Tourangeau，Rips，and Rasinski (2000)，Cambridge University Press授权．Copyright © 2000.

6. 此处文本框 ，来自Schwarz，Hippler，Deutsch，and Strack (1985)，美国公共舆论研究学会授权．Copyright © 1985.

7. 图7.3，来自Jenkins and Dillman in Lyberg，Biemer，Collins，deLeeuw，Dippo，Schwarz，and Trewin (1997)，John Wiley and Sons授权. Copyright © 1997.

8. 此处文本框，来自Oksenberg，Cannell，and Kalton (1991)，瑞典统计局授权. Copyright © 1991.

9. 此处文本框，来自Schuman and Converse（1971），美国公共舆论研究学会授权. Copyright © 1971.

10. 此处文本框，来自Kish（1962），美国统计学会授权. Copyright © 1962.

11. 表 9.2，来自Fowler and Mangione (1990)，Sage Publications授权. Copyright © 1990.

12. 表10.5，来自Campanelli，Thomson，Moon，and Staples in Lyberg，Biemer，Collins，de Leeuw，Dippo，Schwarz，and Trewin (1997)，John Wiley and Sons授权. Copyright © 1997.

目　录

1　调查方法导论

每个月第一个星期五前一天的上午8:30，一组经济学家和统计学家就会走进美国华盛顿特区马萨诸塞大街2号的一个保卫森严、没有窗户的房间。一旦有权进入房间的人到齐，房间就封闭了。

坐在房间里的人都是美国劳工统计局（Bureau of Labor Statisitics）的专家，他们的任务就是审阅和论证对主要经济数据的统计分析。实际上，他们已经花费整周的时间来仔细钻研这些数据，进行比较，检验数据的质量指标，考察奇异值，并撰写描述这些数据的新闻稿。他们用简单的语言撰写新闻稿，目的是让那些没有技术知识的人能理解这些数据是如何产生的。

第二天上午8点，在劳工部大楼隔壁的一间有监测设施的房间，一群记者坐在一起，并切断与外界的一切联系，等候劳工统计局的专家来宣布统计数据。记者们就是根据这些数据来撰写各自的报道的。上午8:30，这些记者们几乎同时要把报道通过电子传输的方式传回自己的新闻机构播发，有时甚至直接采用电话报道方式。

公布的统计数据包括上个月的失业率和创造的就业岗位数。在数据披露前之所以采用如此严密的措施，并使用会被媒体任意演绎的语言来传达，是因为这些数据对社会有着极大的影响。的确，在几个月的时间里，这些数据就是美国经济景气状态的重要信号，无数的股票市场投资者们要据此来决定是买进还是卖出。在消息披露后的45分钟

之内，世界各地的市场都会依据上午8:30公布的两个数据产生数万亿美元的交易。

失业率和就业岗位数据来自统计调查。家户调查生成了失业率，雇主调查生成了就业岗位数。家户和雇主样本都是仔细选择的，综合他们的回答，就能反映所有人的状况。在这两项调查中，成千上万的个体，认真地回答了精心设计的问题，说明自己或公司的状况。在家户调查中，专业访员询问受访者，并将受访者的回答记录到笔记本电脑中。在雇主调查中，雇主使用纸质或电子问卷回答问题。在数据搜集完成以后，还有一个复杂的数据清理和加工过程，用以保证数据的内部一致性。

这两个数据之所以有如此重要的影响，就在于它们是国家经济景气的重要指标，并且是可信的。宏观经济理论和数十年的经验结果都证明了它们的重要性。问题是，只有决策者相信这些数据时，它们才体现出了价值。这本书讨论的就是通过统计调查产生这些数据的过程以及调查设计如何影响统计调查的质量。实际上，问题在于：在什么情况下，调查数据是可信的，什么情况下不可信？

1.1 导言

本章旨在把调查方法作为一个科学的、专业的领域进行介绍。首先将对这个领域进行界定，以便读者在后面纳入应用。在本章结束的时候，读者应该已经理解什么是调查方法以及调查方法有什么用。

一项调查 （survey），是为对更大总体的属性进行定量表述，而针对其中的（样本）群体搜集信息的系统方法。使用“系统”一词是

慎重的，也寓意丰富，是要区分于其他搜集信息的方式。定义中之所以出现“样本”一词，是因为有时候是对总体的每个个体进行调查，有时候则仅仅对样本进行调查。

所谓定量表述，就是统计（statistic）。统计是对一个现象集观察值的定量归纳。有些是描述性统计（descriptive statistic），用于描述总体属性变化的规模和分布，如人们的平均受教育年数、医院就医的总人数、总统的支持率。另一些则是分析性统计（analytic statistic），用于说明两个或多个变量之间有怎样的关联，如用回归系数说明收入与受教育年数之间的关系；教育与过去一年读书本数的相关性。调查的目的使得调查与其他用于描述人和事的方法相区别。统计就是试图描述这个世界大小总体的基本特征或经验。

几乎每个国家都用调查资料来估计失业率，估计接种疫苗对疾病的预防率，估计对中央政府的支持率，估计大选的投票取向，估计对购买物品和服务的满意度等。调查也是监测全球经济趋势、通货膨胀率、对新经济领域投资的核心工具。在社会科学中，调查是理解社会、检验理论最常用的工具之一。在当今信息社会，调查也是关键性的建设工具。

尽管调查包含了一系列活动，但本书只关注具有下列特征的调查：

1）通过向人提问搜集基本信息。

2）通过访员提问并记录受访者应答或通过受访者自访来搜集信息。

3）通过对总体中的群体即样本的访问而不是对所有总体成员的访问搜集信息。

学术（ology）这个来源于希腊文的词汇意味着要钻研（the study of）。**调查方法**（survey methodology）也是关于调查的学术，是关于调查的*误差*（error）来源以及如何尽可能地使搜集到的数据准确的学问。这里的误差是指与期望产出的偏差或距离。在调查中，误差用来指与总体真值之间的偏差。有时用*统计误差*（statistic error）来区别一般意义上的简单差错。

上述的每一步都会影响调查结果的质量（或误差属性），如问什么问题，如何搜集应答，什么人接受访问等。本书将介绍如何在现实世界里进行调查以及如何评估调查结果的质量，说明哪些是已知的，哪些是未知的。本书将尽最大努力把近百年为获得高质量的调查研究所形成的理论、原理、实践展现出来。

1.2　调查研究简史

Coverse（1987）写过一份重要文献，回顾了美国调查研究的历史。这里我们摘其要而述之。有四个方面值得一提：调查的目的，问题设计的发展，抽样方法的发展，数据搜集方法的发展。

1.2.1　调查的目的

最早的调查也许是人口普查 （census），一般由政府执行。人口普查是对人口进行系统的调查，常用于税收或政治代表性等目的。在美国，宪法规定每十年进行一次人口普查，用以依据当前居住格局确定众议院议员的分布。如此，也使得人口普查带有了很强的政治性。也因为如此，人口普查也常常在政治上引起争议（Anderson，1990）。

早期采用调查的一个重要理由是获得对社会问题的理解。有人把现代的调查研究追溯到了Charles Booth，他进行过一项里程碑式的研究《伦敦人口的生活与劳动》（*Life and Labour of the People of London*：1889—1903）（http://booth.lse.ac.uk/）。正如Converse详述的那样，Booth用自己的钱搜集了伦敦大量穷人以及他们为什么穷的数据。根据搜集的数据，他至少写下了17卷著作。但他并没有运用我们今天所运用的方法，即良好的抽样技术、标准化问卷。即使如此，访员们的观察和推论也产生了大量的信息。Booth则通过对系统测量数据的归纳获得了对社会问题的基本理解。

Schuman（1997）关于“民意调查”与“调查”

“民意调查”（poll）与“调查”（survey）有什么不同？民意调查常用于个人观点的研究，运用了很多调查设计的方法。“民意调查”很少用于政府或科学领域的研究。尽管两个术语之间没有严格的区别，但Schuman注意到两个术语有不同的来源，一般认为“民意调查”是德文词，来自于数人头。而“调查”是法文词（survee），最早则来自于拉丁文*super*（超）*videre*（看）。“民意调查”用于反映大众，也是盖洛普（Gallup）、哈里斯

（Harris）和其他民意调查推广的结果。“调查”则是学术界的术语，强调科学的、学者工作的特征。（此处 ）

与研究社会问题相反，记者和市场研究则通过（街访）调查来获得对“普通人”（the man on the street）观点的系统了解。调查的一个特别兴趣是对政治领导人的反应和大选投票取向的了解。这种兴趣推动了现代民意调查的发展。

相对而言，市场研究试图了解人们对现有或未来产品或服务的“真实”反应。早在1930年代，就有过关于电台最受欢迎的节目或信息的调查。研究者采用了更大规模的样本来获得对商业决策更有价值的信息。

在20世纪早期，常常是同一个机构在做民意调查和市场研究，常用的是邮寄问卷调查和电话调查；常用可用的列表进行抽样，如电话号码、驾照、登记选民、杂志订户等。通过对受访者提一些问题，通过访员观察，通过他人代答等方式搜集数据。这些特征使得他们的调查与早先的调查有着明显的区别。与搜集事实和人的客观特征不同，民意调查和市场研究关心的是人们的知识、感觉和想法。

态度和观点测量是现代管理哲学将重点放在消费者满意度上的基础。消费者满意度调查测量购买者对产品或服务质量的预期，以及购买后在多大程度上满足了预期。这样的调查是管理的普遍工具，用于改进组织的绩效。

政治家和政治战略家现在相信民意调查对获得好的竞选决策和了解公众关系的重要议题具有至关重要的作用。的确，对现代政治家的

一个普遍批评就是，他们过于依赖民意调查数据来形塑他们的个人观点，在解决问题时选择公众的观点而不是为大众提供领导者自己的观点。

1.2.2 标准化提问的发展

对测量主观状态的兴趣也导致了对提问遣词造句和搜集方法的关注。当搜集客观信息时，研究者并没有认识到提问时字斟句酌的重要性。通常只是给访员一个清单，如年龄、职业、受教育程度，由访员自己决定如何提问。有经验的研究者常常有着极大的自信，认为自己知道如何提问并获得答案。

但市场研究者和民意调查机构要做大量的访问，需要聘用没有任何社会科学背景的新手做访问。这样，研究者就需要详细说明调查所要获得的信息。不仅如此，研究者还发现，在询问态度访题时，小小的词句变动都会对受访者的回答产生很大影响。

在民意调查的早期，就注意到要给访员经过仔细推敲的访题，并要求每个访员严格依照访题提问。此外，随着访员访问数量的增加，研究者发现访员提问和记录答案的方式对数据也有影响。这些发现最终导致了较之从前对访员更加正式的培训和督导。

在学术界关注商业研究动向的过程中，遣词造句也受到了学术的影响。心理学家对心理状态定量化的心理测量就关注到如何对被试者的心理状态赋值。智力测量就是这个方向上的第一个努力。同时，像Thurstone也探讨过如何对态度、感觉和程度赋值（参见Thurstone and Chave，1929）。

大多数情况下，他们的方法都非常麻烦，基本上也只能找那些学生志愿者来填答那些冗长的问卷。对代表性样本而言，这些工具对大多数调查都不适用，对测量一个或几个态度而言，问卷都太长了。Rensis Likert在他的博士论文中（Likert，1932）使用了一种单个问题加上分程度答案的形式，获得的结果与冗长问卷获得的结果相似。Likert把这种方法运用到调查中，并在1946年创建了密歇根大学调查研究中心。

1.2.3　抽样方法的发展

早期的研究者，像Booth，都是试图去搜集研究总体的所有成员的信息。这种方法避免了因从总体中选择部分进行调查所产生的误差，当然，对大规模总体而言，也不实用。的确，分析普查资料的困难使得人们开始努力将样本资料推及总体。早期的抽样也许是研究一个典型的“镇”，或者有意识地搜集个体资料，使其与总体相似，如访问一半的男性、一半的女性，并使其在地理分布上与总体相似。

尽管18世纪已经有了概率论，直到20世纪概率论才被应用到抽样调查实践中。最早的应用就是从总体中系统地获得“N 中的1”。这就是*概率样本*（probability sample），即每个个体都有被选作样本的非零机会。

抽样应用的最大突破来自于农业研究。为了预测作物产量，统计学家们创造了*面积概率抽样法*（area probability sample），也就是抽取一定面积作为样本来预测农民春季的耕作到秋季会有怎样的收获。人们把同样的方法用在了家户调查中。在城市或乡村，抽选一定

的区域，列出全部的家户，再从家户列表中抽选样本。抽样时，需要找到一种方法让每个家户甚至家户中的人都有被抽中的机会。这种技术的魅力在于，抽样时无须将总体中的所有家户都列出来。

大萧条和第二次世界大战是调查研究的助推剂。现代概率抽样之一就是1939年12月开始的“月度失业调查”（Monthly Survey of Unemployment），29岁的统计学家Morris Hansen主持了这项调查，后来他也成为了这个领域的领军人物（Hansen，Hurwitz，and Madow，1953）。战争期间，联邦政府希望通过调查来了解人们的态度和观点，如购买战争国债的兴趣，以及其他的事实。战争期间，不少资源都用于了调查。在战争期间从事调查的人们，后来对调查方法的发展都起到了重要的作用。战争结束后，方法专家认识到，要想获得基于总体的好的统计，需要关注调查方法的三个方面：问题是如何设计的；数据是怎么搜集的，包括对访员的培训；样本是如何抽取的。

评价其他样本的基础是概率样本。概率样本广泛应用于政府统计机构为政策制定者提供的重要依据中，概率样本也被应用在司法诉讼中。对媒体受众的规模估计也运用概率样本，并由此确定广告率。简而言之，当样本负载大量价值时，通常使用概率样本。

1.2.4　数据搜集方法的发展

早期的资料搜集和尽可能与更多人讨论一些论题只有点滴区别：访员要记录并用统计方法进行归纳。随着搜集系统数据方法的发展，成本在下降，周期变得更短，调查作为工具也变得越来越流行。

对识字的总体而言，邮寄纸版问卷提供了非常低廉的测量。1960年人口普查邮寄问卷的正式测试非常成功，1970年普查时，就大规模地采用了邮寄问卷。与访员面访相比，邮寄问卷要便宜得多。当然，邮寄问卷尽管很好，但受邮路的影响，调查周期则要几个月，而不是几周。

随着电话的普及，市场研究者首先看到了运用媒介来搜集数据的两个优势。电话比邮件快多了，电话比面访便宜多了。但在几十年中，电话调查也深受折磨，无法覆盖没有电话的穷人和总是处于流动状态的人口。但到1990年代，几乎所有的市场调查都不再做面访，不少学术研究似乎也紧随其后，要放弃面访。但是，联邦政府机构却继续进行家户面访。

和许多的人类行为领域一样，计算机的发明给调查的效率带来了本质的跃迁。在美国最早制造的计算机中，就有一台用于每十年一次的人口普查。调查研究者很快就认识到计算机可以大量减少用于调查的人力资源。调查研究者们首先将计算机应用于分析，而后用于数据清理，再后用于编码，直至应用于搜集数据。现在，计算机（从手持机到网络）几乎应用于调查研究的每个环节：调查设计、数据搜集、数据分析。在这个领域成长最快的是基于网站的调查。

随着各领域的发展，调查研究也形成了一些操作指南。经验研究证明了抽样设计对统计质量的价值。访员培训指南促进了访员的标准化，计算和报告应答率的标准使得不同调查测量之间得以比较。

自各种调查诞生60年以来，在通过设计来改善调查统计质量方面取得了长足的进步。但是，正如从这段简短的历史中可以看到的，20世纪的上半叶只是界定了好的调查方法的基本要素。

1.3　几个正在实施的调查的例子

理解调查方法的范围和用调查获得潜在信息的方法就是举例。下面列举了6个例子，整本书中都将因为一些理由不断地提到这些例子。第一，这些都是正在进行中的调查，也都是年复一年进行的调查，受访者认为对他们提供的数据存在着不断的需求。换句话说，有人认为他们是重要的。第二，这些调查并不是特别典型的调查，没有包括民意、政治或市场研究；没有包括任何一次性调查。这些调查都是全国性的调查，也都是受政府经费支持的调查。

当然，这些调查在许多方面都不相同。我们选择这些调查项目的一个理由是：它们覆盖了调查研究的主要主题，运用了多样化的研究设计，为调查研究提供了很好的案例。正因为如此，借助这些案例，我们可以看到调查中众多的方法问题及其解决途径。

下面将要介绍每个案例的基本特征，包括：

1）目的；

2）调查覆盖的总体；

3）抽样资料来源；

4）抽样设计；

5）访员的使用；

6）数据搜集的模式；

7）计算机应用。

读者要从两个方面思考这些问题。第一，把它们看作信息源，我们可以从中学到什么，回答什么样的问题，以及提供了怎样的政策和决策信息？简而言之，为什么要做这样的调查？第二，通过比较设计特点看清楚什么设计能够满足、达成怎样的目标？

1.3.1　全国刑事犯罪受害者调查

美国的刑事犯罪状况如何？刑事犯罪的频率是在上升还是在下降？谁是刑事犯罪的受害者？每个社会都在试图回答这些问题。在美国，是通过对有组织犯罪的公众调查来回答这些问题的。在1930年代，国际警察总长协会（International Association of Chiefs of Police）就开始搜集行政记录数据，根据各类警察个人和办公室所掌握的行政数据汇集成全国的刑事犯罪数据。警长们设计的记录要求有刑事犯罪现场、受害者、犯罪嫌疑人以及任何与犯罪证据相关的资料，通过个人的记录汇集成行政记录。问题是，只有在有人报警时，刑事犯罪才会引起警方的注意。但把报案界定为案件或事故常常由基层警察自己决定。在很多年里，这些记录是美国关于刑事犯罪的关键信息来源。

后来，这些记录在统计上的缺陷逐步显现出来。有时候，一个新市长希望减少犯罪，但却创造了一个环境让警察们更多地将报案界定为事故而不是犯罪，这样，就减少了刑事犯罪的行政记录。此外，不同管辖机构常根据自己的定义对这些记录进行归类，也污损了这些记录。当警察们相信刑事犯罪永远都不可能消除时，他们就会鼓励市民

们不填写正式的报案文件。越来越多的证据显示涉及管辖权的警民关系极差。公众对警察的害怕导致了公众回避报案，而警察自己的态度也影响了其子群体，有的记录，有的则不记录刑事犯罪。越来越清楚的是，尽管主要的刑事犯罪（如谋杀）在行政记录中是具有代表性的，但行政记录越来越忽视了轻微刑事犯罪，常常是不报。一些管辖机构保留着非常详尽的记录，而另一些则只有粗线条的记录。

因此，在几十年里，当人们询问“美国的刑事犯罪情况到底如何”时，就越来越不再相信统计数据的价值了。此外，对刑事犯罪的简单计数不能为政策制定者提供清晰的信息（如犯罪特征、受害者），用以制定减少刑事犯罪的政策。约翰逊总统建立的“总统司法执法与行政委员会”（President Commission on Law-Enforcement and the Administration）注意到，“如果我们知道罪犯和受害者的更多信息，知道他们之间的关系以及犯罪高发的环境，那么防止刑事犯罪项目的效率就会大大提高”（President's Commission，1967，转引自Rand and Rennison，2002）。

1960年代后期，犯罪学家探索一种可能性，即直接询问人们是否是刑事犯罪受害者。由此形成了关于刑事犯罪调查的另一个视角。与关注事件不同，这种探索关注了事件的一方——受害者。这种视角的转换与警察系统的行政记录形成了鲜明的对照。当然，杀人犯罪的受害者是无法受访的。孩童的父母不一定很好地报告孩童受害者（因为他们不知道在学校发生了什么），而孩童自己又不会报告。针对同伴的刑事犯罪，也显示出“好报告”的一些问题。举例而言，如果有人纵火焚烧一个公寓，那么谁是受害者？公寓所有者？承租者？还是失火时的访客？从一个角度看，这些人都是受害者，但如果询问所有人就会将一宗刑事犯罪的计算复杂化。此外，受害者有时还会报告不愉

快的事件（如无权进入他们房子的人进入了他们的房子）。还有，他们也不能像警察那样搜集潜在罪犯的信息（如有人试图偷电视机）。

另一方面，调查受害者能报告那些在警察行政记录中没有报告的案件，由此应该提供了关于刑事犯罪的更全面的情况。的确，从刑事犯罪受害者的角度来看，受害者自报是对警察记录的重要补充。此外，受害者调查的另一个好处就是可以在全国范围内使用标准化的受害测量工具。

但是，进行全国性的刑事犯罪受害测量又面临另一些难题。尽管可以让全美的警察根据行政记录报告刑事犯罪情况，但却没有财力让所有美国人报告刑事犯罪受害情况。只有“代表性”样本，对受害者的调查才有可能。这样，与记录数据比较，调查要面对的是抽样误差（sampling error，即由于缺失了总体中的一部分人而在统计上产生的误差）。

问题是，人们真的能准确地报告他们的受害情况吗？调查方法专家们在1960年代的后期和1970年代的早期研究了报告问题。在方法研究中，研究者针对警察的记录抽样，然后用这些样本去找受害者。他们发现，受害者的报告与记录是相互印证的。但是，Gottfredson和Hindelang（1977）以及其他的人注意到一个问题，就是事件的时间对不上。大多数重要事件的报告时间比实际发生的时间要更接近于当下。但无论如何，受害者报告模式有效地调整了在全国建立完全独立的刑事犯罪检测系统的努力。

当在调查方法中运用所有的设计特征时，有一点会非常清楚，即警察报告和受害者报告的差异就会不可避免（Rand and Rennison，2002）。调查的优势在于，它能测量到警察报告测量不到或警察不记

录的内容。但是，调查却有抽样误差。另一方面，警察记录的犯罪包括一些非本国公民，这类人会在美国的家户调查中缺失。还有，在警察记录中包括有杀人和放火，但却不包括简单的偷袭。警察记录中排除了对男性的强奸，而调查则可包括两个性别的强奸。有些刑事犯罪有多重受害者，每个受害者都会报告相同的状况（如家户犯罪、集团盗窃）；调查可能将其计算为多次刑事犯罪，而警察则将其计算为一次刑事犯罪。警察报告依赖于联邦政府和成千上万辖区的自愿合作，而每个辖区的合作方式又不相同；同样，运用传统方法的调查则会因为无法纳入无家可归者或迁移者而头痛。调查中，如果罪犯很有名，受害者则倾向于少报。同样有证据显示，调查也会少报不大重要的案子，因为受害人不大容易记忆。最后，两种记录都会因为同样的特征而出现重复记录问题。在调查中，同类受害人的报告会被记录为一次“系列案件”（例如，丈夫多次殴打妻子）；而警察记录的则会有人报告一次，记录一次。全国刑事犯罪受害者调查（The National Crime Victimization Survey，NCVS）（表1.1）1972年发端于美国司法部。现在的司法统计局负责搜集和发布犯罪、罪犯、受害者以及各级政府司法运行的信息。司法统计局把NCVS搜集信息的工作外包给了美国人口普查局。

表1.1　调查例子：全国刑事犯罪受害者调查（NCVS）

负责方	美国司法统计局
执行方	美国人口普查局
目的	主要目标是： • 搜集受害者和犯罪后果的详细信息 • 估计没有报告给警察的犯罪数量和类型 • 对某些类型的犯罪提供统一的测量 • 用于历时的和地区间的比较
开始年	1973 年，前期叫"全国刑事犯罪调查"（National Crime Survey，1973—1992）
目标总体	12 岁或以上的青少年、成年人、普通社会成员和非机构成员
抽样框	美国家户，分县、街区、列表地址、列表家户成员
抽样设计	多阶段分层整群区域概率抽样，每 3 年进行抽样单位轮换
样本规模	约 41 800 户（78 600 人）
访员	督导下的访员
访问方式	面访和电话访问
计算机辅助	70%的纸笔问卷访问，包括面访和电话访问；30%的计算机辅助访问
报告单位	每家户 12 岁或以上成员自我报告
时间维度	同地址的接续追踪调查
频率	每月搜集
每轮调查的访问	3 年内每 6 个月对样本家户进行一次访问
观察层级	犯罪事件的受害者、家户
网址	http://www.ojp.usdoj.gov/bjs/cvict.htm

NCVS调查要求人们报告在过去6个月中的刑事犯罪受害经历。如果要求人们报告过去12个月的刑事犯罪受害经历，研究者就可以从每个访问中获得更多的事件，且更有效地利用访问时间。但是，早期的研究表明，如果要求人们报告6个月以前发生的事情，从准确度来看，对事件的丢失率会显著增加。实际上，即使只询问过去6个月内发生的事情，也会少报。如果只询问1～2个月内的情况，准确度会大大上升；如果只询问1～2周的情况，准确度就更高。问题是，随着报告事件的缩短，就会有越来越多的人无事可报，通过访问能够提供的受害信息也会越来越少。设计者选择6个月，实际上是在准确度和访问生产率之间的一个平衡。

NCVS的样本采用阶段接续法抽取，以保证每个12岁或以上的人都有被抽中的机会，进而保证对美国适龄人口的代表性（参见第4章 的多阶段分层整群区域概率抽样）。样本严格限定为家户成员，不包括无家可归者、机构性成员、团体性成员（调查方法专家认为，覆盖这类子群体的成本会极高，以至于会干扰调查目标）。样本整群分布在几百个不同的样本区域（通常是县或一组县），样本设计重复了在这些区域多年进行的研究。采用整群样本是为了便于招聘、培训访员和督导，也便于访员在样本家户之外访问样本家户的成员，从而节省成本。同样出于节省成本的考虑，设计了让每家户12岁或以上的成员接受访问，这样就可以得到更多的受访个体。

这项调查节省成本的另一个方法就是重复访问同一个地址。当设计上随机获得了一个NCVS的样本家户后，访员会询问，除了第一次访问以外，样本家户是否愿意在6个月以后再次接受访问，且继续下去，在3年之内接受7次访问。除了节省成本，这样的调查可对历时的受害率进行较高质量的估计。这种设计被称为轮换追踪设计 （rotating panel design），因为每个月都有受访者接受第一次、第二次、第三次、第四次、第五次、第六次、第七次访问；这样，每个月样本都会轮换，且每6个月轮回一次（译者注：即第1个月的受访对象到第7个月就会被访第二次；第2个月的受访对象到第8个月就会被访第2次，以此类推）。

每一年，NCVS访问42 000家户，包括76 000人。其中，92%的家户受访1～2次，受访家户中87%的个人符合受访者资格，并接受了一次访问。2006年，也是距本书出版最近的一次调查，91%的受访家户中86%的个体接受了访问。

询问的问题包括了前6个月内刑事犯罪受害的频率、特征，以及给家户带来的后果。访问覆盖了家户受害和个人受害类型：强奸、性侵犯、抢劫、侵扰、盗窃、入室抢劫、车辆被盗等。访员到访受访户，询问家户内的居住人口谁在过去的6个月内有遭受刑事犯罪侵害的经历。户内的一个人回答所有财产类刑事犯罪的问题（如入室抢劫、暴力等）；每个个人则报告自己遭受的刑事犯罪侵害（如个人财物的盗抢、侵扰等）。对受访户的首次访问，访员要面访，后面的追访则可以从两个呼叫中心采用电话访问。如此，NCVS的数据搜集是个电话访问（60%）和面访（40%）结合的混合模式。

访问要求受访者报告过去6个月内发生的所有刑事犯罪受害事件，例如文本框中列举的盗窃。

我会读出一些例子，以便让您了解本项研究覆盖刑事犯罪的范围。

我读的时候，如果其中有任何事情在过去的6个月中发生在您身上，请您告诉我该事件具体发生在20______年______月______日。

有哪些属于您的财物被盗？

1）携带的财物，如行李、钱包、手包、肩包、书

2）衣服、首饰、计算器

3）自行车或体育器材

4）家里的财物，如电视、音响、工具

5）车里的财物，如行包、杂物、相机、磁带

如果受访者回答“是的，有人偷了我的自行车”，这时，访员就会先记下这件事，并稍后询问该案的详细内容。图1.1展示了NCVS可计算的统计数，即1994—2000年受访家户遭受3类刑事犯罪之一或以上的百分数。注意，百分比在下降，说明1990年代后期的刑事犯罪率在减少。政策制定者在密切地关注这类数据，并将其作为刑事犯罪治理政策效果的重要间接证据。但研究者却注意到，当国家经济繁荣、失业率低的时候，刑事犯罪率也会下降。

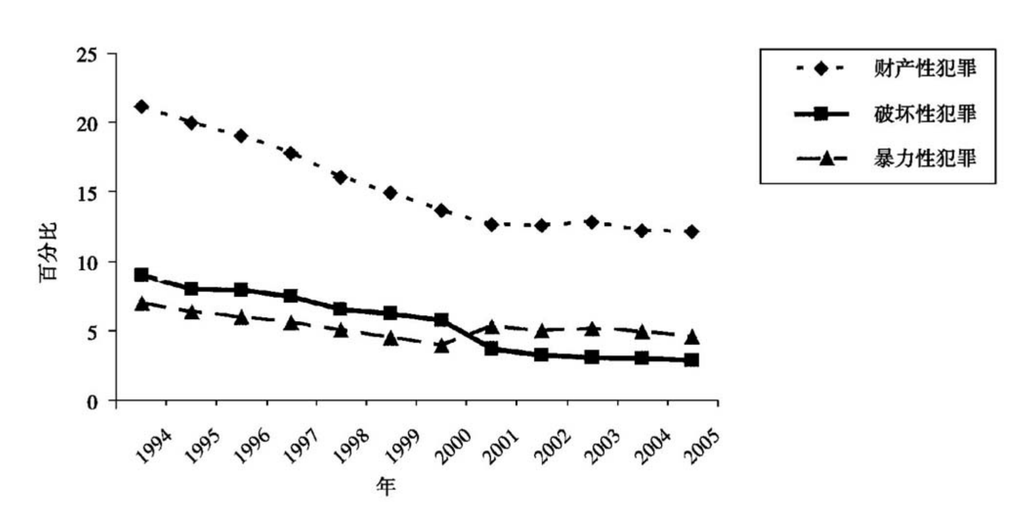


图1.1　美国家户遭遇刑事犯罪的类型，1994—2005年全国刑事犯罪受害者调查

（数据来源：www.ojp.usdoj.gov/bjs）

1998年12月27日CNN.com的新闻标题称，“政府说，刑事犯罪状况达到25年来的低点”。新闻的第一行说，“星期日，美国司法部报告说，去年，美国人成为1973年以来遭受刑事犯罪侵害最少的一年”。接下来的报道说，“克林顿总统对星期日报告的刑事犯罪现状表示骄傲。总统宣称，司法部的数据说明，‘加强警力、实行更加严格的枪支管制，更好地预防刑事犯罪’的策略是有效的。”这就是从NCVS获得的发现。总统把刑事犯罪下降与现行的政策关联在一起是正常的，

但却没有有力的经验支持。（在现有信息下，把刑事犯罪下降与政策执行联系起来是非常困难的。）

1.3.2　全国药物使用与健康调查

在美国，有多大比例的人使用非法药物？在这些人中，穷人和受教育程度低的人比其他人的比例更高吗？随着年龄增长，药物使用有变化吗？药物使用模式随时间在变化吗？不同群体使用不同的药物吗？酒精类滥用与药物类滥用有关联吗？各州的用药模式不同吗？全国药物使用与健康调查（National Survey on Drug Use and Health，NSDUH）每年从各州抽取家户样本。访员到访每个样本家户并访问每个样本个体，询问他们的背景和非敏感性问题。为搜集用药情况，访员为受访者提供笔记本电脑，受访者戴上耳机，收听事先录制好的提问，使用键盘记录自己的应答。每年，NSDUH都要估计几种不同药物的使用率。这些数据用于美国联邦政府的药物政策，目的是为了减少非法药物的需求和供应。

表1.2　调查例子：全国药物使用与健康调查（NSDUH）

负责方	物质滥用与精神健康服务局(SAMHSA)
执行方	RTI(Research Triangle Institute)国际
目的	主要目标是: • 提供使用率、人数的估计值和其他与非法药物、酒精、烟草相关的在州、联邦层次的测量 • 改善国家对物质滥用的理解 • 测量国家减少物质滥用的进展
开始年	1971 年,前期叫“全国药物滥用家户调查”(National Household Survey on Drug Abuse)
目标总体	12 岁或以上的美国非机构成员
抽样框	美国家户,分县、街区、列表地址、列表家户成员
抽样设计	分州的多阶段分层整群区域概率抽样,每 3 年进行抽样单位轮换
样本规模	141 487 住户单位,67 870 人(2007 年 NSDUH)
访员	访员访问,加上敏感问题自访
访问方式	在受访者家里面访,加上部分自访
计算机辅助	计算机辅助面访(CAPI),加上语音计算机辅助下的自访(ACASI)
报告单位	每个家户 12 岁或以上成员自我报告 受访家户可让知情家庭成员完成整个的“健康保险和收入”部分
时间维度	重复性截面调查
频率	每年一次
每轮调查的访问	1 次
观察层级	个体、家户
网址	http://www.samsha.gov/

第一次通过家户调查方式来试图测量药物使用情况是在1971年。估计到年龄与药物使用之间高度相关，调查中更多抽取了12—34岁样本，以便做龄组分析。在调查之前，曾担心美国人接受药物使用调查的意愿，由此，设计者还担心应答率。为此，访员自己还联系样本家户并询问样本个体。当访员询问到涉及药物使用的敏感问题时，就将面访转换为自访，访员还特别说明将为数据保密，让受访者将应答填写在专门的答题纸上，在应答完成之后，将答题纸装入事先准备的信封密封，随访员一起投入邮箱寄走。

最初的设计包括了不少流行的设计理念。像NCVS一样，NSDUH的目标是家户成员和非机构（如难民营、收容所、集体宿舍等）成员，覆盖有居住单元的美国公民中12岁或以上的人员。随着时间的推移，样

本量在不断增大，最近又为50个州各自独立推论进行了重新设计，每年的样本量达到了70 000人。这就使得8个大州每年都能进行独立推断，其他州也能结合过去的资料进行推断。新的设计对青少年也加大了样本量，每个州的样本在3个年龄组（12～17岁、18～25岁、26岁及以上）上都有一样的分布。与NCVS不同，这项调查的样本个体只受访一次，每年都要重新抽样，每年的推断也依据当年的样本，这就是所谓的*重复截面设计*（repeated cross-section design）。大约91%的家户接受了过滤性调查（测量家户的结构），79%的样本个体接受了完整的访问。

调查尽管是由联邦政府的机构即物质滥用与精神健康服务局（Substance Abuse and Mental Health Services Administration）在负责，但执行却外包给了RTI国际。对大型调查而言，这种情况在美国虽然常见，但在其他国家却不是。在历史发展过程中，美国调查研究组织中的私人机构（包括营利性、非营利性、学术机构）与联邦政府和州政府发展了一种合作关系，让大量政府需要的信息由非政府工作人员搜集、加工和分析。

这项调查覆盖了大量的药物（例如酒精、处方药）。设计者不断更新测量程序，目的是获得更加准确的药物使用的自报信息。现在，访员使用笔记本电脑展示访题并记录应答（这个过程叫作计算机辅助面访，Computer-Assisted Personal Interviewing，CAPI）。同时，还进行语音计算机辅助下的自访（Audio Computer-Assisted Self-Interviewing，ACASI），即受访者头戴耳机，通过计算机播放的语音，运用键盘填答问卷。运用ACASI相对面访而言，增加了药物使用的自报量，例如，自报使用可卡因的人数增加了2.3倍（Turner，

Lessler，and Gfroerer，1992，p. 299）。这项调查对处理和预防药物滥用起到了持续的指导作用。

通过每年自报的同类药物使用统计，NSDUH可以提供历年药物使用情况的净变化。但大量的方法研究表明，NSDUH的自报数据倾向于低报。但人们期望，采用相同的方法对历时数据进行比较，如果低报的趋势是不变的，那么历年的变化就应该是准确的。例如，图1.2说明，1999—2001年使用大麻、精神类药物和致幻剂的有少量增加。如此的变化，会导致对旨在较少药物供应和处理药物滥用项目的重新评价。

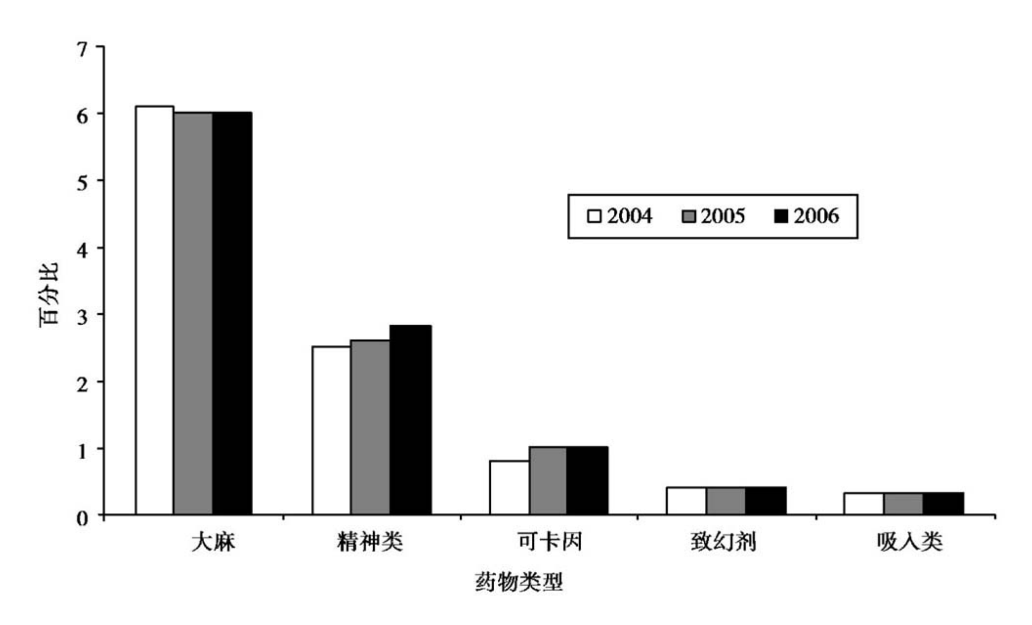


图1.2 过去一个月自报药物滥用情况，2004—2006（数据来源：NSDUH）

2000年2月4日，CNN.com就有一个新闻标题说“克林顿政府要烟草业对青少年负责”。新闻的第一段说，“为了减少弱势群体的烟草消费，克林顿总统要求烟草业为每一位年龄在18岁以下的吸烟者支付3 000美元，并计划对每包烟的征税再提高25美分。”随后，新闻又说，“政府说，根据全国家户药物滥用调查的最新数据，现在有410万青少年烟民。”政策制定者将NSDUH的数据当作了当前反毒品政策成功和失败的证据。当数据表明政策失败时，积极的政府就会制定新的政策。

1.3.3　消费者调查

人们对自己未来的财务状况是乐观还是悲观？近期他们有大宗购买（如汽车、冰箱）计划吗？他们认为现在比过去几年过得更好吗？富人比穷人更乐观吗？乐观的程度随时间在变化吗？

1946年，经济学家George Katona发现，询问人们对个人和国家的经济展望可以提供有关国家经济未来的有用信息，这一点又植根于人们关于未来的态度会影响其消费购买和储蓄行为的基础性发现。进而言之，个体的购买和储蓄决策是经济健康的重要组成部分。从那以后，密歇根大学调查研究中心就持续进行消费者态度调查。这项调查不仅仅有联邦政府资助（事实上，这类统计量是“美国主要经济指标”的一部分），也有来自美国联储局和一些商业公司等与国家货币政策有关机构的捆绑资助。

每个月，消费者调查（SOC）都要抽一组电话号码样本，通过这些号码找到家户，从每个样本家户选择一位成年人进行调查。目标原本是整个家户。为降低成本，则只选电话登记用户（没有电话的人要么更穷，要么居住在乡村，或是过客，参见第3章）。抽样采用随机数字拨号（random digit dialing）设计，即在某区号内抽取样本调查号码（不一定所有抽中的电话号码都是家户电话号码，有些电话是空号、商用号码、传真或上网号码）。每个月，访问500个有效号码。现在，大约60%样本家户的成人接受了访问。在联系某个样本家户时，访员就尝试访问并在6个月后进行追问。因此，SOC像NCVS一样，属于轮换追踪设计。搜集数据以后，还要根据非家户样本或无应答样本进行统计调整。每个月都运用初访和追访数据计算消费者信心指数。

调查的50道题涉及个人财务、商业环境，以及购买环境。调查运用呼叫中心的计算机辅助电话调查系统，即用台式电脑显示访题，接受应答，并指导访员询问下一个合适的问题。

每个月都要发布消费者情绪指数（Consumer Sentiment Index），用于对国家经济增长的未来进行预测。这个指数，结合其他一系列指标，构成了“美国主要经济指标”（Index of Leading Economic Indicators），这些指标被经济预测家们用于为宏观经济政策和投资战略提供咨询服务。令人着迷的是，调查中的一些简单问题能够预测像美国经济这样复杂体系的变化。例如，图1.3显示了两个统计量的曲线。一个是对调查访题的应答，“在未来的12个月中，人们外出工作的情况会怎样，你认为失业的人比现在多、一样，还是少？”图1.3比较了基于对这道访题应答形成的指数（参见左轴）和每年实际的失业率（右轴）。消费者期望指数是黑线，每年实际失业率的变化是浅色线。注意，消费者的期望总是能提前预测月度的失业率。这就说明消费者期望中包含了失业率未来变化的预测性信息，而这些信息却没有被其他经济信息捕捉到（Surveys of Consumers，2003）。

表1.3　调查例子：消费者调查（SOC）

调查名称	消费者调查
负责方	密歇根大学
执行方	密歇根大学调查研究中心
目的	主要目标是： • 测量消费者态度和预期的变化 • 理解为什么会产生变化 • 评价这些变化与消费者的储蓄、借贷决策如何相关，抑或只是随意性的改变
开始年	1946 年
目标总体	美国大陆地区（不包括夏威夷和阿拉斯加）非机构成人
抽样框	美国大陆地区的电话家户，运用有效电话区号和交换机域的电话列表
抽样设计	随机数字电话号码列表，随机选择的成人
样本规模	500 位成人
访员	访员主导
访问方式	电话访问
计算机辅助	计算机辅助电话访问（CATI）
报告单位	随机选择的成人
时间维度	两次调查的追踪
频率	每月 1 次
每轮调查的访问	2 次；初访以后的 6 个月，对初访受访者进行追访
观察层级	个体
网址	http://sca.isr.umich.edu/

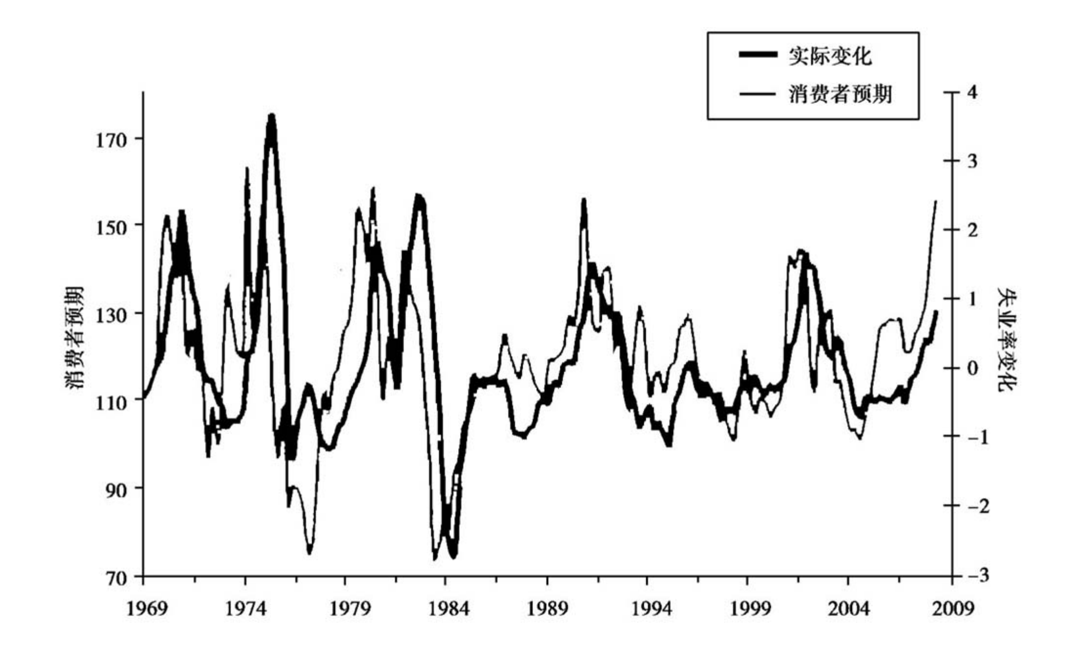


图1.3　消费者失业预期与美国失业率的实际变化，1969—2009（数据来源：消费调查）

投资性企业和股票市场观察着消费者信心统计。他们似乎认为消费者信心统计是很好的预测指标。2002年6月15日，《纽约时报》有篇文章的标题是“下降中的消费者信心促使道琼斯指数下跌”，描述了美国股票交易市场的情绪，放量抛售使得股指短线下跌。为说明信息对股票市场影响的复杂性，同一篇文章列举了巴基斯坦的汽车炸弹影响股指下跌作为另一个例子。有很多信息会影响投资者的决策行为。非常清楚的一点是，其中一项就是消费者信心。

1.3.4　全国教育进展评估

接受初等教育的儿童在数学、阅读、写作上的训练在什么程度？一些学校提供的学习机会比另一些学校更多？来自贫困、少数种族/族裔家庭的儿童比其他儿童的成绩更好还是更差？不同地域或州的孩子在数学和语文训练方面会不同吗？美国与其他国家比又有什么不同？学习成就会历时变化吗？

几十年以来，有研究表明不同学区在不同时期提供的教育是不同的。州教育机构常常把州的教育投入纳入检验，如此，所有州采用的也许是同一个模式。的确，在过去的几十年里，政治人物为了获得家长的支持促进了更多对公立学校的投入，导致了对投入的普遍检验。

遗憾的是，尽管检验很普遍，但却没有统一的评估。每个州，有时甚至是每一个学区，各自采用自己的检验方式。此外，也不是所有的学校都采用了标准化的评估。常常，投入较少的学区就省略了评估

活动。这样，对全国的教育管理而言，对上述问题的回答就不系统了（注意NCVS的情形和警察的犯罪报告）。

1960年代，随着联邦政府在教育中角色的增强，建立统一的教育评估体系就势在必行了（Vinovskis，1998）。不过，通过统计调查提供统一的统计量在政治上是有争议的。这个例子也说明调查统计的重要性甚至会使其本身变成政治问题。

首先，有些州不想与其他州比评估分，因此他们更支持将评估局限在国家层面，而不是在州的层面。其次，人们反对将评估作为工具对每个学生进行评估，因为人们担心用于统计目的的调查被学校、家长、学生在评估时的特别努力侵蚀了，以致不能反映全国的状况。在政治上，也有反对性的争议，与其把钱花在评估上还不如把钱花在改善教育本身上。最后是政治理念上的冲突。一些相信强有力地方控制的人，通过评估似乎看到了联邦将对教育采取的强有力干预。

尽管如此，美国教育部于1969年启动了“全国教育进展评估”（National Assessment of Education，NAEP），用于获得全国性评估资料，1990年又增加了用于州层次的样本（允许对每个州做准确的估计）。NAEP是三个独立评估的集合体：“全国评估”（提供全国年度数据）、“各州评估”（各州的估计）、“趋势”（历时评估）。每个评估包括4个部分：小学和初中学生调查，学校特征和政策调查，教师调查，残疾或英语障碍学生（SD/LEP）调查（用于NEAP主体），或者被排除学生调查（用于NEAP趋势）。在各州评估进行了多年以后，就没再设立独立的全国样本了；全国的样本来自于各州样本的累积。对NAEP的长期趋势而言，采用的是残疾/英语学习（SD/ELL）学生调查。1985年，作为NAEP的一部分，在教育考试服务与责任分析公司的资助下，又开展了青年人识字调查，用于评估21—25岁个体的识字

能力。此外，NAEP还进行了高中成绩单研究。但在本书中，我们将主要关注全国评估。

表1.4　调查例子：全国教育进展评估（NAEP）

负责方	美国教育部国家教育统计中心
执行方	Westat
目的	主要目标是： • 评估4、8、12年级学生在一些主题上的能力 • 反映当前教育和评估的实践 • 测量历时变化
开始年	1969年
目标总体	全国评估：在校4、8、12年级的学生 各州评估：在校4、8年级的学生
抽样框	美国分县、县组的列表学校和列表学校的在校小学生和中学生
抽样设计	分州的多阶段分层整群区域概率初级抽样单位，在初级抽样单位内抽取学校，在学校内抽取学生教室
样本规模	全国评估：2 000所学校，100 000名学生 各州评估：每个州100所学校，每个调查年级2 500名学生
访员	无；校长、教师、学生背景自访；调查执行机构指导下，学生考试
访问方式	纸笔自访问卷和考试卷
计算机辅助	无
报告单位	校长、教师、学生
时间维度	重复性截面调查
频率	每年一次
每轮调查的访问	1次
观察层级	学校、班级、学生
网址	http://nces.ed.gov/nationsreportcard/

国家教育统计中心作为联邦政府的统计机构，负责NAEP，但数据的搜集像NSDUH一样外包出去了。“教育考试服务”是一家公司，负责学术技能测试（SAT）和其他的标准测试，以用于评估。Westat是一个调查机构，设计并选择样本，负责样本学校的调查执行。NCS Pearson也是一家测试与教育评估机构，负责计算评估得分。

全国评估选择50个州和哥伦比亚特区在校的4、8、12年级学生。抽样设计为整群初级抽样单位（县或县组），与NCVS和NSDUH一样，这是为学校行政与评估行政之间的合作而采用的节省成本的办法。在抽选区域以后，再选择学校。在每个样本学校，直接选择相应的年级，使得样本分布在不同的班级。NAEP每年都有不同的主题。此外，在每个样本学校，针对不同的学生要做不同的评估。针对学生的测试内容也做随机区分，但主题是同一个。

每个评估都预先准备了专门的概念框架。由这些概念框架共同界定一组知识，以反映学生能力不同的复杂层次。一套复杂的、长期一致的评估过程，把教师、教育专家、家长、学校行政以及一般公众整合进了一个框架。然后，专家们编写和测试每道访题，用以代表框架中的每个部分。与前面介绍的调查不同，这里是多个访题测量同一个框架中的不同部分。例如，在数学框架内，会有不同类型的数学知识（如4年级就有加减乘除的内容）。

NAEP是全国学校绩效的重要指标。教育资助方运用NAEP的信息调整教育资助的目标和层次。NAEP提供涉及各科教育成就、教学经验、学生总体（如4年级学生）和学生亚总体（如女生、西班牙裔学生）环境等信息。这些都是主要政治和政策非常关注的信息。图1.4呈现了不同类型学校高中高年级数学的平均得分（500为满分）。这个发现符合人们的日常观察：私立学校学生的得分较高，天主教学校居中，公立学校的平均得分最低。总体上，也没有一个一致的变化趋势。例如，2000年，除了天主教学校以外，其他学校的得分都在下降，但总体上都比1990年要高。当出现分数下降时，就会引发公众讨论和政策辩论，涉及导致下降的原因和教育政策的调整。2000年以后私立学校的参与率低于70%的最低值，所以没有报告。

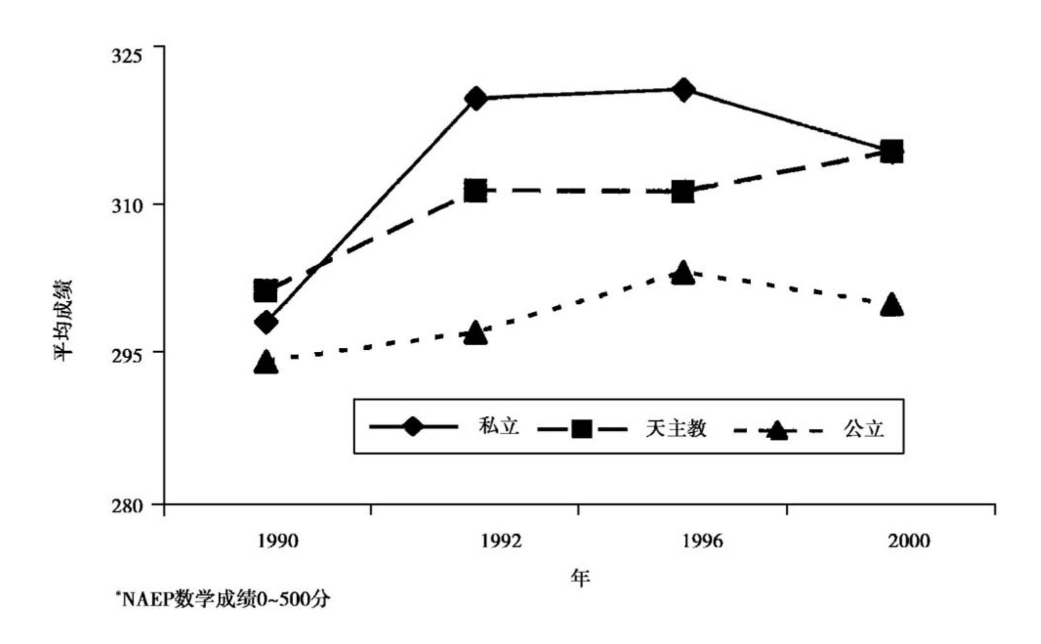


图1.4　分年分学校类型12年级学生数学测试的平均分（数据来源：美国教育部，教育科学研究所，国家教育统计中心，全国教育进展评估，1990，1992，1996和2000年的数学评估）

2003年6月20日CNN.com有一则标题为“小孩比高年级学生更聪明吗？”的报道写道：

> 星期四，政府报告说，4年级的学生表现了更好的阅读能力，与此同时，12年级学生的阅读能力在下降……总体而言，低于1/3的4年级和8年级学生展现了理解和分析具有挑战性资料的能力。这个能力等级，即熟练，是测试的关注点。在高中的高年级学生中，36%的学生达标。就在4年前，29%的4年级学生达到了熟练，2002年增加到31%。那些更年轻的学生成为了国家改善基础教育的中心。在高年级学生中，达到最高等级的比例从40%降了下来……“没有科学答案能够告诉我们高年级学生为什么在阅读测试中表现得如此糟糕，但我们会继续寻找战胜挑战的方案”，教育部Rod

Paige说，“与此同时，我们知道了如何让更年轻的学生提高阅读能力，也知道所有的学生都能学习。”

这些是复杂的结果，需要更多的分析来澄清2002年的12年级与1998年的12年级之间是否存在群体差异，根据经验，这应该是多个项目或者是多种教育政策共同作用的后果。缺乏机构性数据，使得这个结果对政策制定者而言存在多种解决方案。除了评估层次混合因素的复杂性以外，对这种现象的理解还需要NAEP提供统一的测量。

1.3.5　行为风险因素监测系统

有多少人锻炼、抽烟或系安全带？美国各州的这些现象有差异吗？各州与健康相关的行为有怎样的差异？随着年龄变老，人们会有更多还是更少的健康行为？在各州，公共卫生项目的实施是否在时间序列上体现了人们健康行为的变化？

1965年以后，国家卫生统计中心每年都提供多种健康相关行为和状况的调查结果（如自报健康状况，看医生的频率，自报锻炼情况，以及其他行为风险相关因素）。调查数据与生物医学研究相结合，清楚地说明个体行为会影响早期的疾病犯病率和死亡率。许多公共卫生政策和观察都是州级的，但是，州级的统计数据却没有比较。各州的卫生机构在安排资源以减少行为相关的健康风险方面具有主导作用。

1980年代以后，行为风险因素监测系统（Behavioral Risk Factor Surveillance System，BRFSS）与美国疾病控制中心合作提供了州级关键健康因素的结果。与上面介绍的项目不同，BRFSS在联邦疾

病控制中心的帮助下把每个州都作为合作伙伴。各州决定自己的问题并执行调查，联邦疾病控制中心确定一组核心问题并执行数据搜集标准，进行数据清理，并发布合并后的联邦层级的数据。核心问题询问当前健康相关的观念、状态、行为（如健康状况、健康保险、糖尿病、吸烟、部分癌症筛查程序，以及HIV/AIDS风险），以及人口特征。

表1.5　调查例子：行为风险因素监测系统（BRFSS）

负责方	美国疾病控制中心
执行方	各州不同;2007 年的 BRFSS,12 个州或主权地区自己调查,42 个外包
目的	主要目标是: • 搜集统一的、各州需要的预防成人疾病和与慢病、病伤、可预防的流行病相关的风险行为数据 • 各州可以比较,并可在联邦层次做结论 • 识别历时趋势 • 允许各州因地制宜设计问题 • 允许各州用附加模块的形式突出本州急迫和紧急的问题
开始年	1984 年
目标总体	美国成年有电话的家户
抽样框	用电话区号或交换局号列表获得有电话家户,再列出样本家户的成员
抽样设计	2007 年的 BRFSS 各州都采用了概率样本设计,运用非比例分层随机样本,但 Guam,Puerto Rico 以及美国的维尔京群岛都采用了简单随机样本
样本规模	2007 年的 BRFSS 各州平均的样本量为 8 309 人
访员	访员主导
访问方式	电话访问
计算机辅助	54 个地区采用了计算机辅助电话访问,2 个地区采用了纸笔访问
报告单位	随机选择的成人
时间维度	重复性截面调查
频率	每年一次
每轮调查的访问	1 次
观察层级	家户的成人
网址	http://www.cdc.gov/brfss/

与SOC一样，BRFSS运用随机数字拨号抽样方法。但与SOC不同的是，BRFSS让每个州独立抽取样本，并使用计算机辅助电话访问。大多数州都委托商业或学术调查机构搜集数据，部分州自己搜集。各州的样本量也不同，但在每个样本家户，都访问18岁或以上的成人。

调查每年发布各州抽烟和其他风险性健康行为的结果。这些结果作为人口健康状况的社会指标，用于指导政府的影响健康相关行为的政策。图1.5的3幅地图展现了美国1990年代肥胖症的急剧增长。每个州的数据是分别展示的，用不同的阴影展示其之间的差异。1994年，没有一个州肥胖症成人超过该州成人人口的20%，即身体质量指数（BMI）大于等于30。到2001年，超过一半以上的州肥胖症流行。在州级追踪这类趋势，使得州公共卫生机构既可以将本州与其他州比较，也可以对本州做历时比较。各州用BRFSS数据估计和追踪本州的健康目标，制订健康项目，或执行具有更广指向的疾病预防活动。

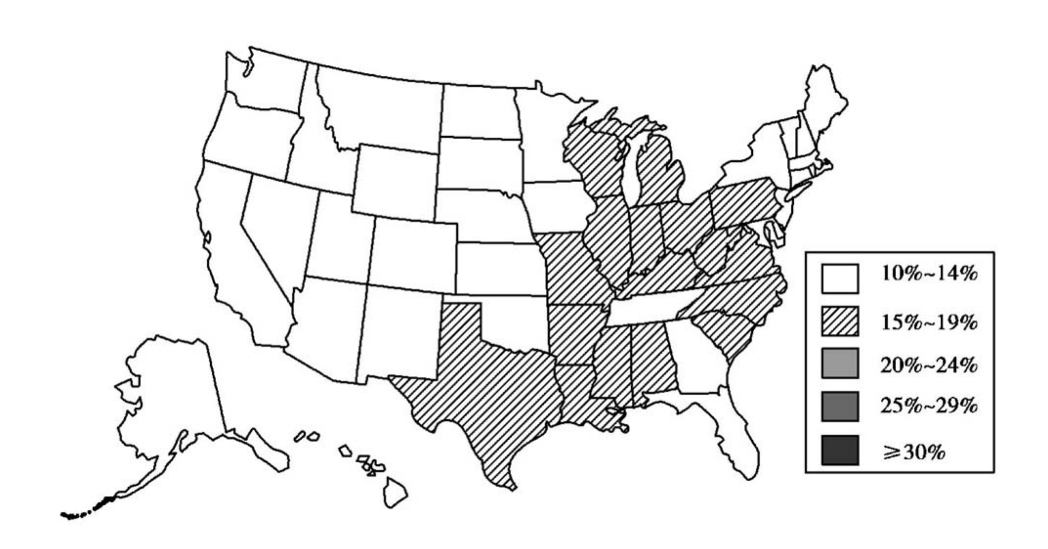


图1.5a　各州肥胖症成人的比例（BMI≥30），BRFSS，1994（数据来源：BRFSS，CDC）

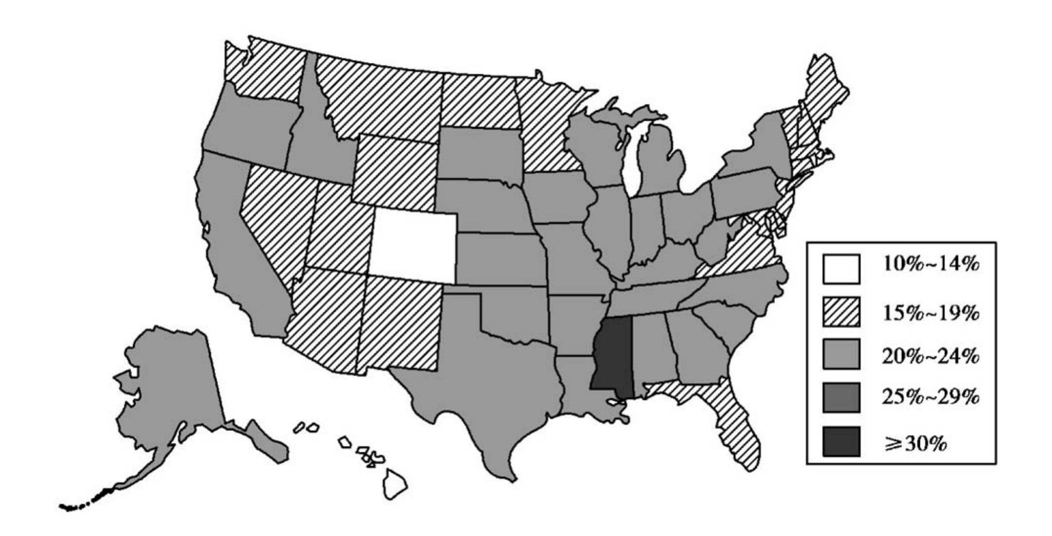


图1.5b　各州肥胖症成人的比例（BMI≥30），BRFSS，2001（数据来源：BRFSS，CDC）

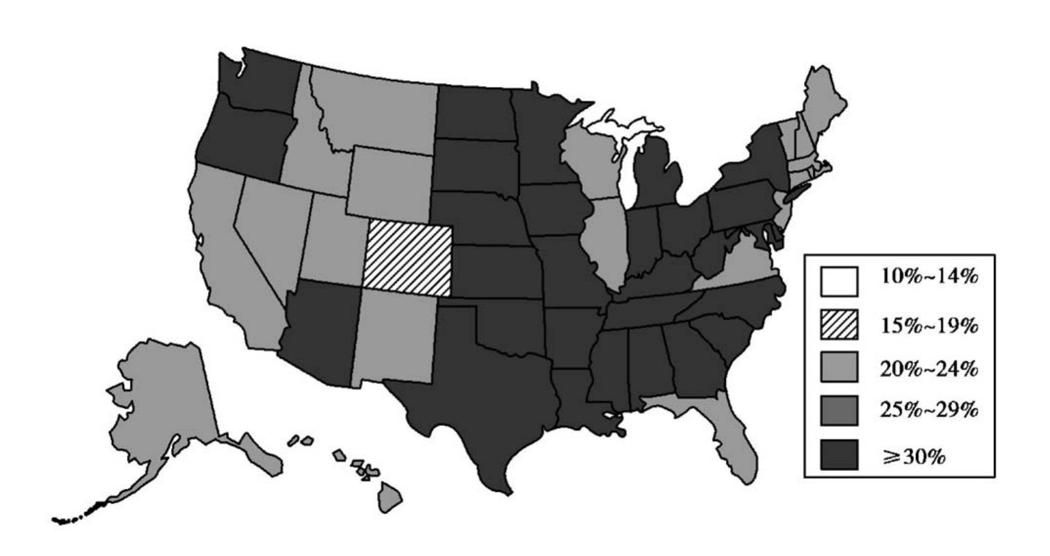


图1.5c　各州肥胖症成人的比例（BMI≥30），BRFSS，2007（数据来源：BRFSS，CDC）

BRFSS数据很快就引起了全国的争论。在美国，尽管有其他指标显示美国人口的肥胖症问题，但BRFSS在地方层次展现了具体的问题。2003年3月6日《华盛顿邮报》有一篇报道说，政府用数据说明，“如果说有明星闪耀，那就是圣安东尼奥——德州的一个城市，成为了美

国肥胖之都”。2003年5月14日，《华盛顿邮报》报道，“肥胖症带来的健康成本与抽烟相当，美国健康与人类服务局（HHS）对快餐业施加压力”。2003年7月2日，《华盛顿邮报》以“瘦身：Kraft让食品更健康”为题报道了美国学校餐厅计划减少分量、选择，改变市场策略的事实。

类似于BRFSS的调查数据，对重要指标的测量结果是有信誉的，也会引起政策决策者的注意。

1.3.6　当前就业统计调查

上个月美国经济创造了多少就业岗位？某些行业比另一些行业的就业数据变化更快吗？哪些行业在上升？哪些行业在下降？变化主要来自于大雇主？还是新的小雇主在就业市场上具有更大的动态？就业的增长或下降有地区差异吗？

美国劳工统计局的当前就业统计调查（Current Employment Statistics Program，CES）就是本章开始介绍的部分。CES是两个平行的用于估计月度就业状况的调查之一。CES是对雇主的调查，询问6个不同的信息：有薪雇员的数量，女性雇员数量，生产工人数量，有薪生产工人数量，生产工人的工时，生产工人加班工时。当前人口调查（Current Population Survey，CPS）则是一项家户调查，询问住户是在职工作还是在找工作，提供月度失业率。

CES的样本量较大，每个月调查超过150 000个雇主。样本来自于在各州失业保险机构登记的雇主列表。抽样设计的策略是，给大雇主高入选概率（部分是永久性样本）、给小雇主低入选概率。雇主一旦

入选，就会每月甚至多年作为样本。小雇主则采用轮换制，大雇主待在样本中的时间要长得多。CES是联邦劳工统计局和各州就业保障机构之间的长期合作项目。

表1.6　调查例子：当前就业统计调查

负责方	美国劳工统计局，美国劳工局
执行方	美国劳工统计局，各州就业保障机构
目的	CES 的主要目标是获得每月联邦、州、大城市的就业、工时、薪酬估计
开始年	1939 年
目标总体	美国雇主
抽样框	在美国就业保障机构进行了失业保险税务登记的雇主
抽样设计	原为配额样本，2002 年完全变成了概率样本
样本规模	约 150 000 企业机构
访员	绝大多数是自访，约 25% 为电话访问
访问方式	不少样本单位（约 10%）通过拨入电话提取语音提示并按提示用电话按键进行音频数据输入（TDE）完成访问，其他的包括邮寄、传真、网站、电子数据交换以及电话访问。
计算机辅助	音频数据录入（TDE），电子数据交换（EDI），网站填答，计算机辅助电话访问（CATI）
报告单位	企业联系人
时间维度	历时追踪雇主调查
频率	每月
每轮调查的访问	1 次
观察层级	企业雇主
网址	http://www.bls.gov/ceshome.htm

与上述的各项调查不同，CES采用多种方法同时搜集数据。雇主有多种方式用于每月提交6个问题的数据。可以用纸版表格填写好之后邮寄或传真给劳工统计局，也可以使用音频数字输入方式，既可用电话的音频拨号方式提供答案，也可以采用加密网站填答方式，还可以使用电子格式的记录回传，当然也可以在接受电话访问时口头回答问题。这些不同的方法适合于不同的雇主，让雇主采用自己适用的方式，有助于促进雇主与调查者之间的合作。

为了说明CES如何追踪美国经济，图1.6显示了1941—2007年非农雇主提供的岗位数。看看几十年的趋势，可以清楚地看到岗位数周期性的停滞、缓慢增长和快速增长。例如，1980年代和1990年代早期的衰退之后是岗位数的快速增长。对2000年代早期衰退的平复也很明显。

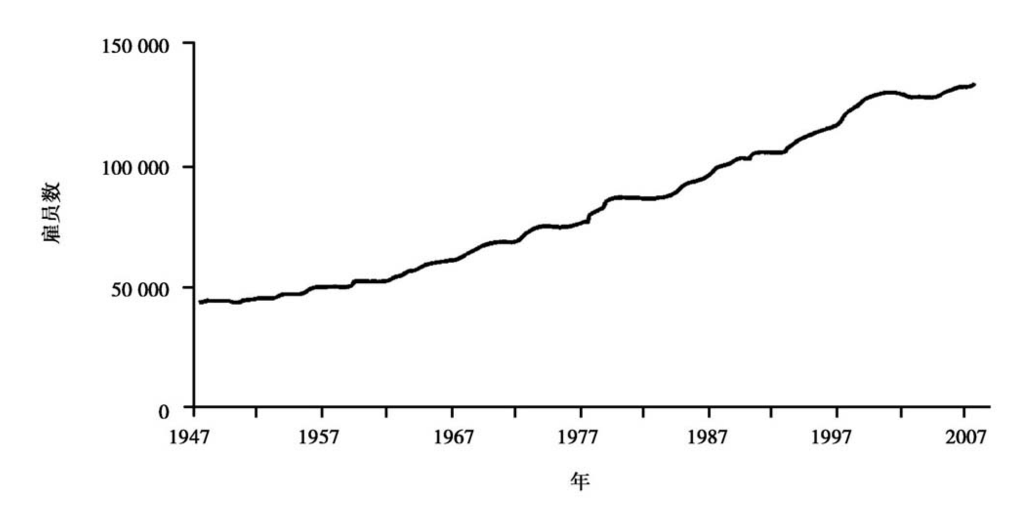


图1.6　所有非农雇主的雇员数（千人），历年估计值（1947—2007），当前就业统计调查

（数据来源：www.data.bls.gov）

这类统计数据非常重要，因为联储局、白宫、国会都要用它来调整经济政策。周期性的衰退总是让政治家们制造说辞，提出必须通过刺激经济来增加就业。不管是否公平，这类统计数据的价值在于它影响了政府声望的起落。

1.3.7　从6个调查例子中我们能学到什么？

所有这些，都是全国性的大规模重复调查。其中，5项调查有悠久的历史，当前就业统计调查早在1939年就开始了，消费者调查则始于1946年，全国教育进展评估开始于1969年，全国刑事犯罪受害者调查和全国药物使用与健康调查也始于1970年代。所有这些调查的数据都被政府机构和社会科学家们用来作为社会或经济状况的指标。

这些调查的特点就是其设计与目标相一致。每一个调查的目标总体都不一样，例如NCVS关注12岁或以上的少儿和成人，而CES的目标总体是雇主群体。NCVS、NSDUH以及NAEP都采用两个阶段来获得家户样本，第一阶段为地理区域抽样，然后在抽中的样本区域产生家户列表。CES和BRFSS则在电话区号或交换局内用所有可能的电话号码抽样，从电话号码中获得家户样本。

这些调查也采用了不同的数据搜集方法。NAEP不使用访员，由受访者自访并评估。NCVS、BRFSS和SOC都依赖于访员去访问并获得答案。NSDUH运用了两种技术，部分由受访者自访，部分由访员访填。这些调查的第二个差别就是对计算机的应用。NAEP是先用纸笔，然后采用计算机加工。2009年，NAEP会进行子样本受访者的计算机辅助访问测试，2011年计划让8年级和12年级的学生在线填答问卷。CES则在多方面采用了计算机辅助，包括音频数据输入、电子数据交换交换，以及计算机辅助电话访问。

所有调查都在继续。这是因为这些调查的设计就是要测量他们所研究对象的变迁，所有调查设计都是全国性的，观察总体历时的均值和总和的变化。对同一受访者搜集资料的次数要大于1次，如CES、SOC和NCVS，就可以测量个体或雇主以及其所代表的总体的变化。

这些设计特征决定了调查的成本和质量。对一个新的调查而言，首先要问的是：

1）目标总体是什么（研究谁）？

2）抽样框是什么（如何确定谁有机会成为调查样本）？

3）抽样设计是什么（如何抽取受访者）？

4）搜集数据的方式是什么（怎样获得数据）？

5）是多次调查还是一次性调查？

请再次阅读1.3.1—1.3.6节，看看每个调查的目的和抽样设计。纵观本书，我们将会用这些调查来解说调查方法的具体问题。

1.4　什么是调查方法？

调查方法是在成本和数据质量约束下，关于调查设计、数据搜集、加工和分析的基本原则，即意味着在给定成本的条件下，改善调查质量；或在给定质量要求的条件下，降低成本。“质量”是用误差参数（将在第2章讨论）来定义的。调查方法既是一门科学，也是一项专业化的工程。

站在调查的科学一面，要获得高质量的数据，就要应用许多传统学科的原理。懂数学，尤其是懂概率论或随机事件，对理解各种产出

的频率是非常重要的。统计学有一个分支，就是研究概率原理和从样本到总体的推论。抽样原理和分析，历史上就是基于数学的。

但是，无论是访员还是受访者，在调查中都会受到多个学科的约束。当涉及资料搜集原则如何影响调查推论时，心理学就是知识的重要来源。此外，社会心理学是理解访问中访员行为对受访者影响的知识框架。当涉及访题设计时，认知心理学对考虑记忆的形成、结构以及什么器物有助于帮助人们回忆起调查访题的答案提供了重要的知识基础。社会学和人类学提供了社会分层和文化多样性的知识，让我们了解涉及某些群体针对特定测量或问题的反应性特征。计算机科学提供了数据库设计、文件加工、数据安全、人机界面等相关的知识。

由于调查涉及了众多学科的交叉，直到现在才逐步发展成为一个系统的领域。调查方法科学面的实践，是不在传统学科领域划分之内的。例如，把概率抽样理论的重要发展应用于调查是从20世纪30—40年代大型政府调查组织开始的。抽样的早期文本是科学家为特定的调查环境如美国人口普查局写的（例如Hansen，Hurwitz，and Madow，*Sample Survey Methods and Theory.* 1953；Deming，*Some Theory of Sampling*，1950）。许多涉及调查数据搜集的重要成果都是来自政府机构在第二次世界大战期间进行的军队调查、公民调查。早期的主要文本，同样都是科学调查机构的科学家结合调查实践中出现的问题撰写的（Kahn and Cannell，*The Dynamics of Interviewing*，1958；Hyman，*Interviewing in Social Research*，1954）。

一以贯之的情形是，研究文献或调查方法始终散杂。统计学期刊关注调查方法，如*Journal of the American Statistical Association*（Application section），*Journal of Official Statistics*，以及*Survey Methodology*。涉及调查研究方法报告的

重要成果是*Processing of the Survey Research Methods Section of the American Statistical Association* 。此外，跨学科的，如*Public Opinion Quarterly* ，也包括其他学科的或应用领域如涉及健康、犯罪、教育、市场研究的，也是刊登调查方法文章的主要期刊。

在这个领域，也有活跃的专业组织，把科学家和专家整合到一起。在调查方法领域，至少有4个广受关注的组织。美国统计学会的调查研究方法专业委员会的会员人数最多。美国公众舆论研究学会（其国际性相关组织为世界公众舆论研究学会）则包括商业性、学术性以及政府的调查研究者。国际调查统计学会是国际统计机构的一部分。商业性的调查组织也建立了自己的协会，美国调查研究组织理事会，目的在于促进成员的利益发展。所有这些组织，包括美国统计学会的调查研究方法专业委员会，都是跨学科的，有来自各个学科背景的成员，如数学的、社会的或不同的应用学科。近期，一个新的组织成立了，那就是欧洲调查研究学会。

调查研究也是一个职业，专注于设计、搜集、加工、分析调查数据。世界各地的学术机构、政府、商业组织都有调查方法的专业人士。在美国，学术机构用调查方法的调查研究者涉及了社会学、政治学、公共卫生、传媒研究、心理学、犯罪学、经济学、交通研究、老年学以及其他学科。在美国的大学校园，有超过100家的调查研究中心，那里有全职的工作人员，为学校的教师提供调查服务。在调查数据搜集方面，联邦政府比学术机构甚至起到了更加重要的作用。人口普查局、农业部、劳动力统计局，都在搜集数据，不仅如此，还有超过60个机构在搜集定量数据。此外，美国私立的商业机构在数据搜集方面的努力远远超过联邦政府各机构之和，如投票、政治调查、市场研究等。

正因为实地调查方法不是在单一的学科内发展起来的，历史上相关的教育活动也很分散。相关的职业也没有职业证书，尽管对职业证书的讨论不断。不过，涉及调查和市场研究技术的培训已经存在一些年了，几乎所有社会科学和各类职业培训中都有涉及调查方法的课程。对于专注于调查方法的人而言，在调查机构当学徒是培训不可省略的一部分。解决设计和执行调查中问题的实践经验是正式课堂培训的重要组成部分。因此，本书的一个重要目的就是为理解和强调调查研究中的问题提供知识基础。

1.5　调查方法面对的挑战

调查并不是搜集大量数据的唯一途径，也不一定是“正确”的途径。政府和商业机构的行政记录系统可以为数据覆盖范围好的决策提供必要的数据。定性调查，包括民族志或社会学的定性调查，可以为对相关主题的纵深理解提供丰富的资料。对人们行为的观察也可以产生关于公共空间事件频率的量化信息。随机实验具有可控环境，可以提供“刺激”与“反应”之间关系的重要证据。

但是，在行政记录系统，研究者对测量没有任何掌控机会。大量的行政记录也许因为数据质量不高而失去意义。民族志调查常常运用知情人小组讨论，也因此缺乏对大规模人群的描述能力。一些人的观察对人类行为整体而言常显得零碎。随机实验不得不面对真实世界的挑战。

同样，调查也限于针对大量人群的标准化的、可重复的测量。正因为调查是在非控制性的现实世界中展开的，也难免受到其影响。调查的推论性来自于其对总体中的子群体（样本）的测量能力，也因此

很难达到完美。调查方法的部分任务就是为追求完美而涉及调查数以千万计的个体的大量决策。其中重要的决策包括：

1）如何确定样本以及如何选择样本？

2）用什么方法联系到样本？在难以接触和不愿接受访问的受访者身上花多大的努力搜集数据？

3）花多大的努力评估问卷和测试问卷？

4）如何针对受访者提问并记录回答？

5）如果有访员，花多大的努力培训访员和进行访问督导？

6）花多大努力检查数据的准确性和内部一致性？

7）用什么方法评判调查估计值以修正发现的误差？

上述的每一个决策都会影响到从调查获得估计值的质量。通常，但不总是，决策都涉及相应的成本；如果决定花更大的努力以获得更大机会去减少误差，常常会伴随着要花更多的经费。

一些方法文献已经说明了涉及数据质量的各种努力。本书的一个重要目标，就是把现在所知的呈现出来；对未知的，也试图说明这些决策如何会影响数据的质量和可信度。

本书的第二个目标是试图让大家理解总调查误差概念及其应用。非常重要的是，所有调查都会存在对理想状态的某种妥协。有些妥协是因为成本。研究者必须决定针对上述的问题如何安排资源。在优先

预算的前提下，研究者常常会侧重于调查的某些方面，并在另一些方面减少成本。譬如，研究者可能决定增加样本量来放弃旨在提高应答率的努力。

调查方法面对的一个挑战就是如何最优地利用可用的资源，即如何在每个分项上分配资源来使数据的价值最大化。在认识到调查的每个方面都会潜在地影响调查结果的前提下，调查方法要考虑的就是总的误差，不是某几个因素，而是作为整体的所有因素。一项调查，重要的不是某个方面比其他方面好，而是在设计和执行中，什么方面最差。总调查误差方法意味着从广义视角保证调查的设计和执行的某个方面不至于差到影响到调查的目标。

同理，在某些情境下，也会存在解决调查设计问题的不完美方式。所有可用的方式在解决具体问题时或多或少都会有些不完美。调查方法学家必须在不完美的解决方法中，确定什么是最优的选择。重复一遍，总调查误差方法就是要考虑不同的方法选择对结果数据质量的影响，并通过平衡，选择能获得最优数据的方法。

调查方法就是关于合理地进行权衡、决策，并最大限度地理解决策涉及的方方面面的知识。一位训练有素的方法学家用总调查误差方法作出这些决策时，总会考虑到决策涉及的方方面面及其对最终结果的影响。

1.6　关于本书

本书描述调查对结果质量影响的具体内容。第2章阐述调查中“误差”的含义。事实上，这是一个模糊的议题，因为不止一种误差会影

响到调查的各种估计。不同的学科也使用不同的术语来表述相同的或相似的概念。后续的每个章节都会关注如何在调查中使误差最小化。理解误差的含义对于理解后续章节的内容至关重要。

后续章节涉及的都是调查设计和执行的各个方面。因此，也涉及研究者的决策，即为什么这些内容是重要的？既有的知识能够告诉我们相关方面的可用选择有哪些？所谓的最佳选择常常涉及调查的目标和其他设计特征。极少情况下有清晰的最佳实践，正在使用的调查方法总会不断产生这样或那样涉及数据质量的问题。后续的章节试图让读者了解决策选择的方式。在整本书中，我们把具体的实践案例（前面所述的6项调查）和一般性知识和原则结合起来，这样，读者可以理解具体决策的获得方式。

最后两章与其他章节有所不同。第11章，我们讨论调查研究涉及的伦理问题。最后一章，我们回答了调查方法学家常被问到的一组问题。当读完本书的时候，我们期望读者通过阅读和课程作业等获得坚实的基础，如果要成为调查方法学家，则还需要实践经验。

关键词

语音计算机辅助下的自访（ACASI）

分析性统计（analytic statistic）

面积概率抽样法（area probability sample）

计算机辅助面访（CAPI）

普查（census）

描述性统计（descriptive statistic）

误差（error）

概率抽样（probability sample）

随机数字拨号（random-digit dialing）

重复截面设计（repeated cross-section design）

轮换追踪设计（rotating panel design）

抽样误差（sampling error）

统计（statistic）

统计误差（statistical error）

调查（survey）

调查方法（survey methodology）

进一步阅读资料

全国刑事犯罪受害者调查

Pastore, Ann L. and Maguire, Kathleen (eds.) (2008), *Sourcebook of Criminal Justice Statistics* [Online].

Available: *http://www.albany.edu/sourcebook/.*

Taylor, Bruce M. and Rand, Michael R., “The National Crime Victimization Survey Redesign: New Understandings of Victimization Dynamics and Measurement,” Paper prepared for presentation at the 1995 American Statistical Association Annual Meeting, August 13-17, 1995 in Orlando,
Florida（http://www.ojp.usdoj.gov/bjs/ncvsrd96.txt）.

全国药物使用与健康调查

Gfroerer, J., Eyerman, J., and Chromy, J. (eds.) (2002), *Redesigning an Ongoing National Household Survey : Methodological Issues* , DHHS Publication No. SMA 03-3768. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Statistics.

Turner, C. F, Lessler, J. T, and Gfroerer, J. C. (1992), *Survey Measurement of Drug Use* , Washington, DC: National Institute on Drug Abuse.

消费者调查

Curtin, Richard T. (2003), *Surveys of Consumers: Sample Design* , *http://www.sea.isr.umich.edu/.*

Curtin, Richard T. (2003), *Surveys of Consumers: Survey Description* , *http://www.sca.isr.umich.edu/.*

全国教育进展评估

U.S. Department of Education. National Center for Education Statistics. *NCES Handbook of Survey Methods* , NCES 2003-603, by Lori Thurgood, Elizabeth Walter, George Carter, Susan Henn, Gary Huang, Daniel Nooter, Wray Smith, R. William Cash, and Sameena Salvucci. Project Officers, Marilyn Seastrom, Tai Phan, and Michael Cohen. Washington, DC: 2003.

Vinovskis, Maris A. (1998), *Overseeing the Nation's Report Card* : *The Creation and Evolution oj National Assessment Governing Board* , Washington, DC: U.S. Department of Education.

行为风险因素监测系统

Centers for Disease Control (2005), *BRFSS User's Guide* , *http://www.cdc.gov/brfss/pdf/userguide.pdf.*

Centers for Disease Control (2008), *BRFSS Questionnaires* , *http://www.cdc.gov/brfss/questionnaires/pdff/ques/* 2008*brfss.pdf.*

当前就业统计调查

U. S. Bureau of Labor Statistics (2003), *BLS Handbook of Methods* , *http://www.bls.gov/opub/hom/home.htm.*

U. S. Bureau of Labor Statistics, *Monthly Labor Review*, *http://www.bls.gov/opub/mlr/mlrhome.htm.* (Website contains frequent articles of relevance to CES.)

作业

---

1. 登录“全国刑事犯罪受害者调查”的网站（http://www.ojp.usdoj.gov/，查看司法统计局），找到“全国刑事犯罪受害者调查”的问卷，判断家户问卷调查搜集的是哪类受害者数据；判断自访问卷搜集的是哪类犯罪数据。

2. 登录“行为风险因素监测系统”网站（http://www.cdc.gov/brfss/），完成提供给访员的培训内容。

3. 阅读最新的“全国药物使用与健康调查”（http://www.oas.samsha.gov/nhsha.htm）年报，说明近些年流行的药物。

4. 找到“当前就业统计调查”的技术说明（http://www.bls.gov/web/cestnl.htm）。说明从不同渠道获得数据的受访者分布状态。

5. 从“消费者调查”网站（http://www.sca.isr.umich.edu）找到最新消费者信心指数，然后到新闻网站搜索“消费者信心”。比较新闻对报告数据的处理，说明数据报告与新闻使用的异同。

6. 登录“全国教育进展调查”网站（http://nces.ed.gov/nationsreportcard/），找到最新的评估

报告，阅读并与过往的报告进行比较。试图找到对观察到的变化的解释。作者从调查数据中引用数据还是援引外部非调查数据的说法？

7. 想一想在阅读本章之前你知道的调查。回头看1.5节列出的问题，对照问题，根据你知道的那项调查进行回答。如果某个具体问题没有答案，写下你是在哪儿找答案的。

8. 重复截面设计和轮换追踪设计都涉及不同时点的测量。两类设计有什么不同？每种设计的什么估计值在不同时点是可以变化的？

9. “全国刑事犯罪受害者调查”和“消费者调查”的目标总体和抽样框如何能不同？对每项调查而言，抽样框和目标总体的差异是什么？又如何影响到每项调查的关键统计量？

2　调查中的推论与误差

2.1　导言

调查方法就是试图理解在调查统计中为什么会产生误差。从第3章到第11章，会详细说明测量误差及使误差最小化的策略。为便于欣赏那些章节并理解调查方法，首先需要透彻地理解“误差”的含义。

让我们从通过调查对总体进行统计描述入手。图2.1是一个示意图。左边是调查的原始资料，如个体对问卷的应答，这些原始资料的价值在于其很好地获得了受访者的特征信息（左上方）。调查对单个个体的特征信息从来没有兴趣，有兴趣的是统计值（statistics），即把个体、群体的应答归纳出来。抽样调查就是把所有个体受访者的应答综合起来，通过统计计算（图2.1中间的云图部分）来建构用于描述样本中所有个体的统计值。在这个意义上，调查只是达成目标（描述样本所代表的总体特征）的步骤之一。

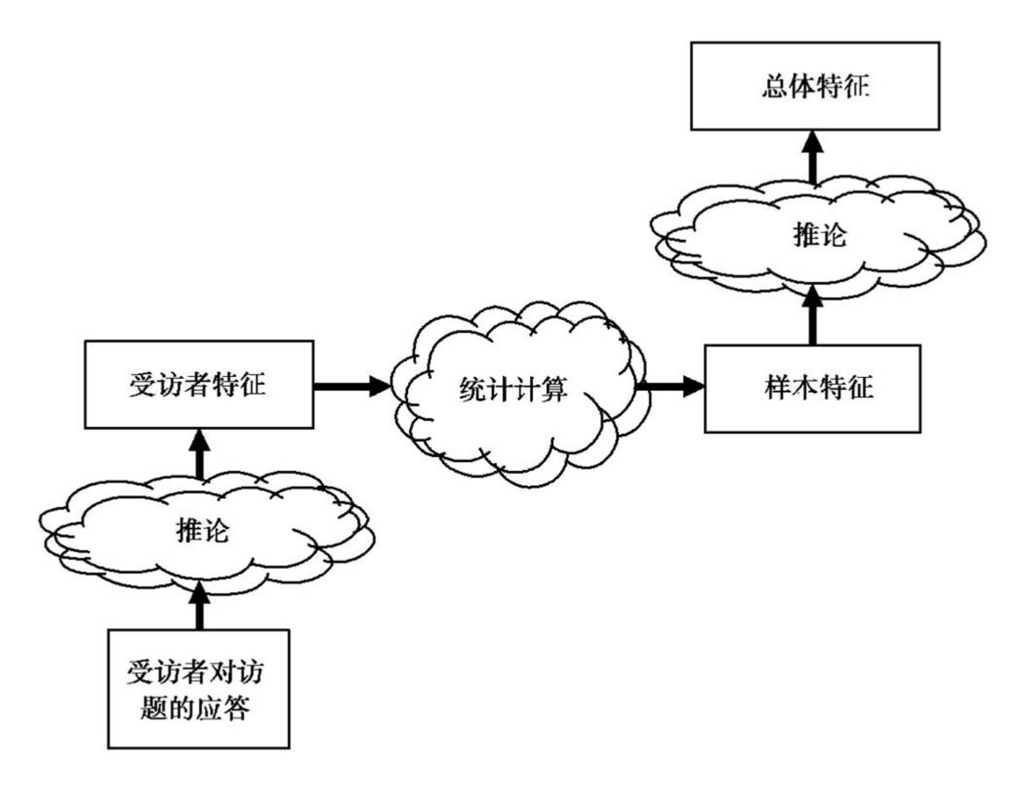


图2.1　两类调查推论

图2.1的纵向箭头标识“推论步骤”，即用获得的不完全信息去描述更加抽象的、更大的总体。在调查中，推论 （inference）是一套正式逻辑，即用观察到的现象来描述未观察到的现象。例如，推论未观察到的思想状态如观点，就是依据受访者对涉及观点的具体问题的应答来实现的。对总体中未测量要素的推论依据是对从同一总体中获得样本的观察。在调查方法中，我们运用受访者对访题的应答来推论受访者的特征，并通过统计值的计算来推论更大总体的特征。

上述两阶段推论是调查两特征的核心：

1）受访者的应答必须准确描述受访者的特征。

2）调查中，受访者子集必须与更大的总体具有相似的特征。

当上述两个条件的任何一个得不到满足时，调查统计值就会有*误差*（error）。这里所说的误差不是日常生活中的错误，而是指与调查过程中所期待的统计值之间的偏差。*测量误差*（measurement error）或*观察误差*（error of observation）是指受访者的应答与期待测量的属性之间的偏差。*非观察误差*（error of non-observation）则指基于样本的统计值与总体之间的偏差。

让我们举一个实在的例子。“当前就业统计调查”（CES）项目旨在测量某个月美国的职位总数。项目询问样本雇主某月第12天那周其支付薪水的雇用人数，由此就会产生误差，因为调查并不询问一月中其他周的雇用人数。有些雇主的数据不完全或不及时，由此也会因为记录差造成误差。这样，用应答值来推论期待测量的特征，就会产生问题（图2.1左边的纵向箭头）。

抽中的雇主样本来自于调查前月份政府发放失业救济金的雇主清单，这样，新产生的雇主就不在其列。因此使用过时的雇主清单就会产生误差。从雇主总体中选择样本雇主，不一定反映了总体全貌，由此就会产生抽样误差。还有，不一定所有的雇主都应答，这样也会因应答缺失而产生误差。如此，用应答统计值来推论总体统计值，也会有问题。

读到此，人们的第一个反应是，各种误差似乎使得调查不可能成为描述大量总体的有用工具。不要绝望！除了这些潜在的误差来源以外，事实上，仔细设计、执行、分析的调查是获得信息来描述这个世

界的唯一工具。调查方法关注的正是影响调查统计值所含信息多少的因素。

调查方法将CES例子中的误差分为不同类型。对每一种误差都有相关的文献，这是因为每一种误差都有不同的影响因素，且对调查统计值有不同的影响。

学习调查的一个方法就是研究每一种误差，或者从“质量”的视角出发。“质量”视角是调查方法特有的视角。学习调查的另一种方法是关注建构一项调查所必需的所有调查设计决策，如识别研究的合适总体（即选择一种方法将总体列表），选择抽样策略，选择数据搜集方式等。这也是大多数教科书的方法（如Babbie，1990；Fowler，2001）。

2.2　从设计开始，一项调查的生命历程

在本节和下一节，我们将介绍调查领域的两个主流视角：设计视角和质量视角。本节讨论设计视角，即如何把一个抽象的想法变成具体的行动。质量视角关注的则是从影响统计值质量的误差来源出发来看设计。我们还是先看设计视角。

调查就是从设计到执行的过程。没有一项好的设计，就不要奢谈获得好的调查统计值。随着关注点从设计转移到执行，工作的重心也从抽象转移到具体。图2.2展示了调查两个平行的方面：测量的建构和总体属性的描述。这个图是从图2.1详析而来。测量讲的是针对样本观察单位 （observational unit）所要搜集的数据：调查什么？而代表

性维度关注的调查描述的总体，即关于谁的调查？这两个维度都需要事前想好、计划好，并细致地执行。

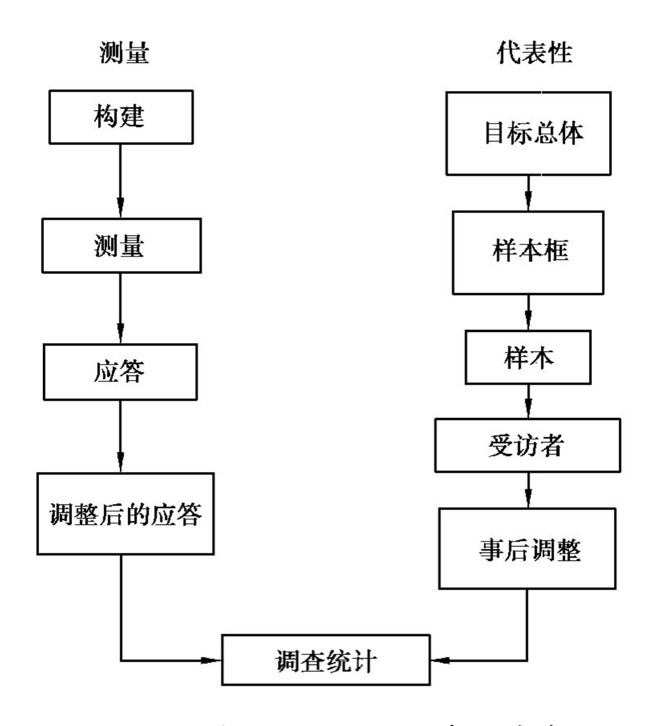


图2.2　从设计开始，一项调查的生命历程

鉴于图2.2涉及了调查方法的重要要素，我们需要多花些时间来讨论，对每个部分进行界定和举例。

2.2.1　构建

构建　（contruct）是研究者搜寻的信息要素。当前就业统计调查（CES）要测量的是过去一个月美国创造了多少个就业岗位，全国教育进展评估（NAEP）测量的是在校生的数学知识，全国刑事犯罪受害者调查（NCVS）测量的是过去一年有多少刑事犯罪的受害者。需要调查的是“多少”，看起来很简单。其实，这样的表述是不准确的，也相对抽象。“多少”没有准确地表述要测量的内容，也没有精确表述测

量到的构建。在有些情形下，构建是理想化的，多数情况下甚至只能用修饰词来表述。

例如，对刑事犯罪受害者的识别就有歧义。当家户内发生暴力行为，譬如把邮箱震掉了，谁是受害者？在这种情况下，NCVS区分家户受害而不是个人受害。如果在公共空间涂鸦，谁又是受害者？是不是“受害者”仅指那些够得上被起诉的犯罪活动的受害者？在什么条件下，让人不愉快的事情才可以被认为是犯罪？所有这些都是人们从用形容词表述到测量操作必须要想的问题。一些构建比另一些更容易转化为测量。

一些构建则比另一些构建更加抽象。消费者调查（SOC）的是个人财务状况的短期乐观性，测量的实际是个体的态度，是不能直接观察到的。态度是个人内在的，个体自己和个体之间都有广泛的差异性。例如，仔细关注自己财务状况的人也许回答得很好；而不关注自己财务状况的人，也许就胡诌一通了。相反，全国药物使用与健康调查（NSDUH）测量上个月的啤酒消费，这个测量就比较接近观察到的实际行为，且可替代的测量不多。唯一的问题是要确定什么样的饮品算是啤酒？例如无醇啤酒算不算？还有用什么作为计量单位，罐还是瓶？消费者乐观性构建就比啤酒消费构建要抽象。

2.2.2 测量

测量 （measurement）比构建更明确。调查中，“测量”就是获得构建信息的方法。调查测量的方式非常多样化：从受访户院子里采集的土壤样本用于获取有毒物质的含量，健康调查中的血压，访员对

受访户住房结构的观察，在交通状况调查中使用电子测量工具对交通流量的计数。不过，调查中的测量常常是摆在受访者面前的访题，而访题是用词语组合的，例如，在过去的6个月，您是否打电话给警察，报告发生在您身上且您认为是犯罪的事儿？测量最关键的就是设计访题，使其答案完美地反映我们试图要测量的构建。用于提问的访题可以是能念出来的（如电访或面访），也可以是可阅读的（如纸笔或计算机辅助的自访）。有时候还包括访员观察，例如让访员观察样本居住单元的结构类型，或观察样本户的邻居。有时候会使用电子或器具测量，例如样本商店物品价格的电子记录，在健康相关调查中采血或采集毛发样本，在有毒废弃物调查中采集土壤样本，采集涂料样本。有时候还会在让受访者看了一段视频以后再提问，例如给受访者看笔记本电脑上的商业视频展示、杂志封面等。

2.2.3　应答

调查中产生的数据来自于通过调查测量获得的信息。应答（response）的性质常常是由测量的性质决定的。当拿提问作为测量工具时，受访者就可依据访题设计的应答选项的变异性来作答。受访者就依据自己的记忆或判断来选择答案，例如“现在向前看，你认为在未来的一年你的财务状况会更好？或更糟？或与现在相似？”（来自SOC）。人们也可以通过查看记录来作答。例如通过检视雇主人事档案来报告第12周有多少非管理职位的雇员（来自CES）。人们也可以让其他人来回答问题，例如让配偶来回答受访者最近一次看医生的时间。

有时候，应答实际是访题的一部分，受访者要做的就是选择其认为符合的类别。另一些时候，调查只提问，受访者要用自己的语言来作答。有时候，受访者也会不提供访题的答案。这就使得涉及该测量的统计计算复杂化了。

2.2.4　调整后的应答

在一些数据搜集中，初始测量来自于对先前测量的评估。在计算机辅助测量中，对数值类应答要做域值监测，并对可接受范围之外的应答进行标记。例如，如果提问出生年，应答为1890的话，就超出了可接受范围，并要通过追问进行校验。此外，还应该有一致性检验，即两个测量之间的逻辑一致性。例如，如果受访者回答她14岁，生了5个孩子，就需要追问并根据追问来修改错误的应答。在有访员提问的纸笔访问中，常常要访员回看已经问过的访题，检查不合逻辑的应答，以及被跳过的访题。

在所有受访者应答完毕后，有时需要对数据进行调整，涉及对应答分布、奇异应答模式等的检查。这种奇异值检测（outlier detection）甚至会指向对某些特定访问问卷的检测。

调整应答的目的是改善原始的应答。调整后的应答就可用于针对构建的推论。

2.2.5　目标总体

现在我们可以来看看图2.2的右边，从抽象转入调查中涉及代表性属性的具体问题了。第一个框涉及的是目标总体（target population）概念，即要研究的单位的集合。如图2.2所示，这是最抽象的总体。对许多美国的家户调查而言，目标总体可能就是“美国的成年人”。这个表述没有对时间进行限定，例如2004年活着的人；也没有说明是否包含了在传统家户之外居住的人，也不清楚是否包括那些刚刚成年的人，更不知道什么居住身份才算。在某些情况下，说明不具体对讨论影响不大；但在另一些情况下，也会有较大的影响。目标总体是一个有限规模的个体集，也是要研究的总体。NCVS的目标总体就是年龄在12岁或以上、处在非兵役状态、居住在非集体性住处的人口，集体性住处如医院、监狱、学生宿舍不包括在其中。时间点为抽样时的月份。

2.2.6　样本框总体

样本框总体是指在目标总体中有机会被选为样本的集。简单的状况是，样本框等于目标总体所有单位（如人和雇主）的列表。但有时候样本框是目标总体的不完全单位集。例如，SOC的目标总体是美国成年家户总体，但却用电话号码列表作为其样本框总体，并将每个人与其家户的电话号码关联。注意，这里的情况比较复杂，有的人没有家户电话，有的人可能有几部家户电话。NSDUH则把美国的县级地图作为样本框，并把每个家户与特定的县进行关联，进而将目标总体中的12岁或以上的儿童和成年个体与其生活的居住单元联系起来。注意，这里的情况也比较复杂，有的人没有固定住处，有的人有多个住处。

推论总体和目标总体

通常，调查统计被用于描述不那么容易被测量的总体。例如，SOC就是要估计在具体月份美国成年人中消费者的敏感性。每一分钟，都会有因结婚或分租组成的家户，也有因死亡、离婚、迁徙导致的家户解体，还有家户的合并等。每月的月初与月尾，家户总体总不相同。有时候“推论总体”指一个月中任何时间里符合资格的个体集。如果样本框是在月初设定的且与该家户的联系也在该月，则“目标总体”是指可被纳入的部分。

2.2.7　样本

从样本框（sampling frame）获取的调查单位就是样本。样本就是调查测量要针对的对象。在许多情况下，样本是样本框的一小部分，也是目标总体的一小部分。

2.2.8　应答值

几乎在所有调查中，都不可能完成对所有样本的调查。对那些成功应答的对象，我们称之为应答值（respondents），于此，也有无应答值（non-respondents）或样本无应答（unit non-response）。如果样本对象只是提供了一部分信息。在确定为应答值或无应答值时也会有一些困难。在制作数据文件时，这些问题必须说

明，即是否包括不完全信息的记录。选项缺损值 （item missing data）是指一个样本中只有某些问题没有应答的情形。图2.3说明了调查的类型、样本框数据、单位的性质，以及选项无回答。

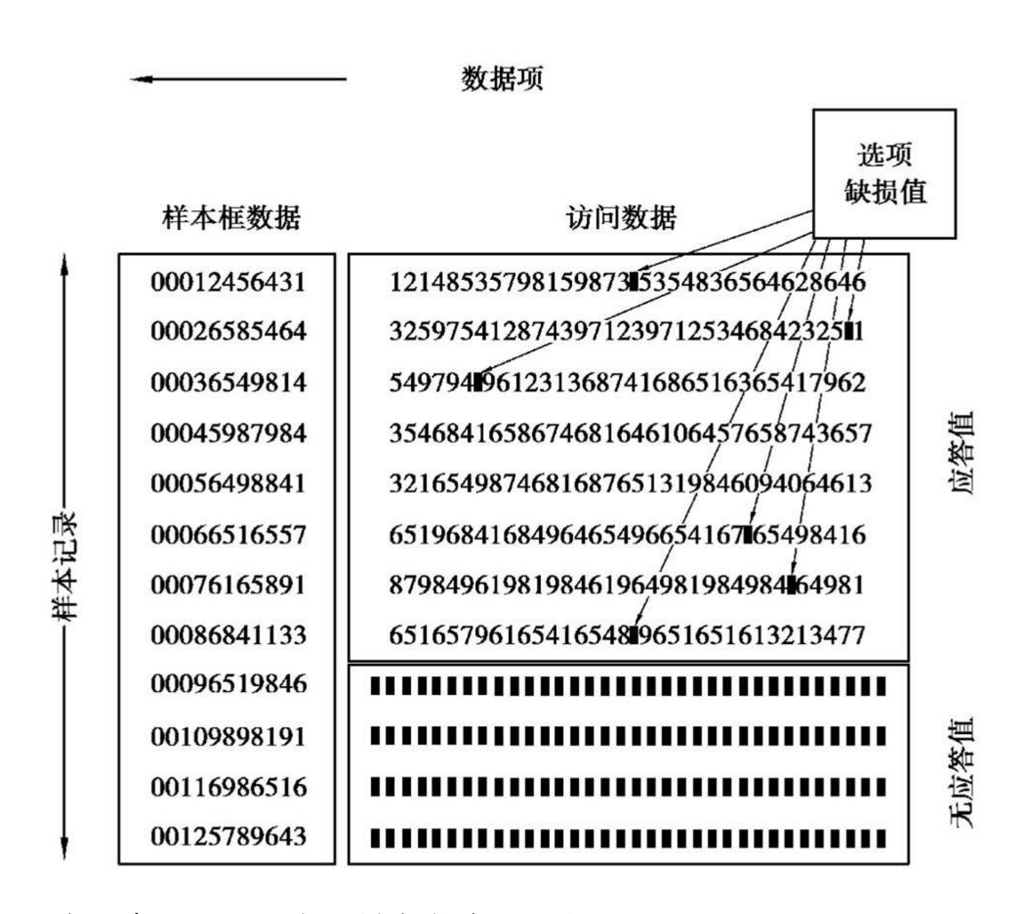


图2.3　一个调查数据文件中的样本或选项无应答

图2.3是一个数据文件的一部分，每一行是一条样本记录，左边是来自样本框的数据，包括了所有样本。应答值数据记录包含了受访者对问题的回答。样本无应答值（文件的后面部分）部分则只有样本框数据。在数据文件中，我们还能看到一些黑块，那就是选项缺损值；在CES中，可能是问卷必须回收时，样本雇主还没有发完薪酬呢。

2.2.9　事后调整

所有受访者提供的数据，就组成了一个数据记录集。为提高根据调查数据进行估计的质量，还有一些工作要做。无应答和覆盖偏差（样本框和目标总体匹配有偏）问题存在，就会对用样本数据估计总体产生影响。有鉴于此，对样本无应答的考察可以了解对样本框中某个子群体（例如城镇的应答率低于乡村）代表性不足的影响。同样的，知道样本框没有包括那些样本类型（例如SOC中的新家户或CES中的新雇主）可以了解其对某类目标总体的代表性不足的影响。后面，我们将要学习对代表性不足的样本进行加权 （weighting）以提高调查估计的准确性。另一个替代方案就是对缺损值进行补值（imputation）。许多加权和补值的过程都被称之为事后调整（postsurvey adjustment）。

2.2.10　从设计到操作

上面讨论设计过程都具有可预计的后果。人们常常按部就班来操作调查，让调查从设计步入操作。

图2.4说明调查的目标如何帮助两项决策，一项涉及样本，另一项涉及测量过程。数据搜集模式 （mode of data collection）的决定对测量工具（例如图2.4中的问卷）的形塑意义重大。在用于搜集数据之前，必须要对问卷测试。在右边的行动中，依据抽样设计进行的抽样框选择将产生用于调查的样本。而在调查阶段，既要用到测量工具，也要用到样本，关注的焦点也变成了获得样本的完整测量（即避免无应答）。在获得数据以后，就要对数据进行调整和编码，使其适用于数据分析。用于分析的数据文件常常包含着事后调整，即针对覆盖性和无应答的调整。调整后的数据方可用于最后的估计或分析阶

段，也就是对全部目标总体进行统计推论的阶段。本书认为，好的调查估计需要同时对调查过程的每个阶段都给予同等的关注。

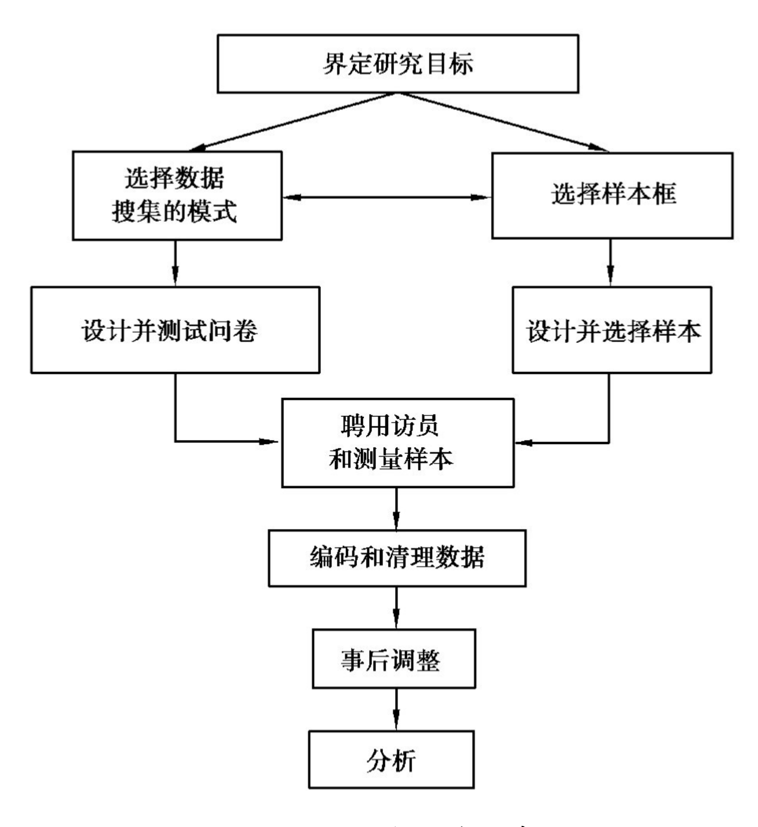


图2.4　从过程看调查

2.3　从质量开始，一项调查的生命历程

我们曾经用图2.2描述了调查的术语。同样的图也可以用来表述调查方法学者对调查质量的思考。图2.5用椭圆框描述了调查方法中涉及的一般质量概念。每个概念都处在调查过程的环节中，用以说明质量问题就出在环节的不匹配上。多数椭圆框中都有“误差”一词，这也是常用的质量术语。调查的设计者就是通过设计和估计环节之间的可

能选择，尽量减少调查统计中的误差。所以，这个框架叫作总调查误差 （total survey error）框架或总调查误差范式。

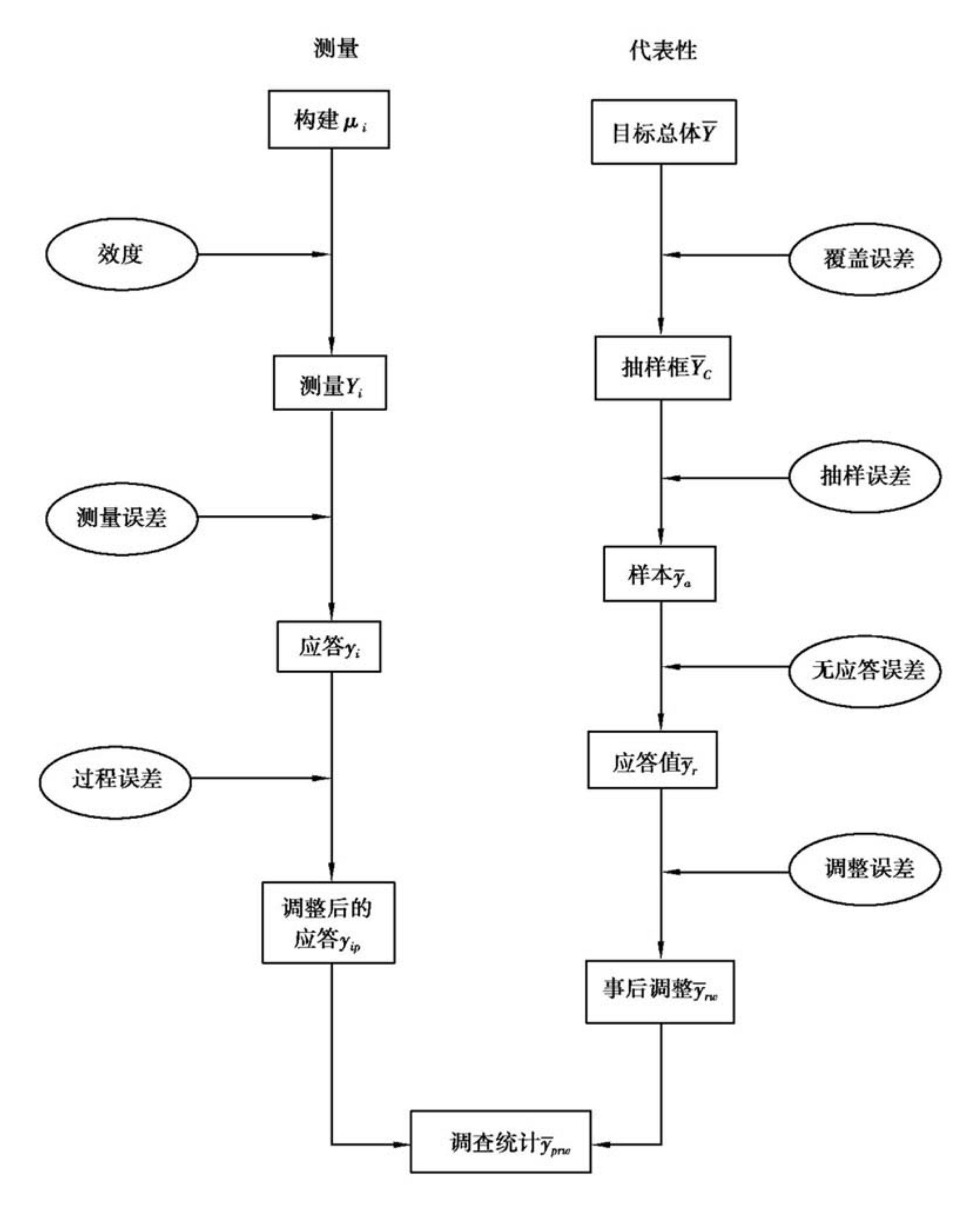


图2.5　从质量开始，一项调查的生命历程

对图2.5，有两点非常重要：

1）每一个质量要素（图2.5的椭圆框）都有文字表述和统计公式。

2）质量要素是单项调查统计值的属性（即单个调查的每个统计值在质量上都可能有差异），而不是整个调查的属性。

下一节将说明各种质量要素概念，并用简单的统计术语进行表述。由于质量不是调查的属性而是单项统计值的属性，因此，可以把这类统计值当作一般统计值的变异（例如样本均值、双变量回归系数、总体估计值）。为使讨论尽可能简单，我们将描述非常简单的统计值的误差项，例如样本均值就是总体均值的一个指标，用于某些建构。样本均值的质量属性，就是其与总体均值之间的函数关系。

我们将用符号来描述误差，这是表述的传统。希腊字母 μ 表示测量目标的不可观察构建。大写字母 Y 表示 μ 的测量方法（对 μ 的测量总会有些问题的）。测量得到的应答用小写字母 y 表示。

这样，统计约定就是：

μ_i =总体中第 i 个人构建的值（如受访者报告的看医生的次数），i =1，2，…，n

Y_i =第 i 个样本的测量值（例如大夫访视数），i =1，2，…，n

y_i =应用测量时获得的应答值（例如对一个调查问题的应答）

y_{ip} =调整或其他过程后的应答值

简言之，我们试图测量的目标属性是 μ_i，但我们实际测量的是不完美指标 Y_i，之所以说其不完美是因为在测量的意义上它偏离了目

标。当用于实际测量时，其结果于我们的理想是有距离的，因为我们实际得到的不是Y_i，而是对测量的应答y_i。我们希望通过调整过程来弥补这类缺陷，调整后的结果就是y_{ip}，即调整后的应答，其中下角标p 指数据获得后。

2.3.1　建构与测量之间的差距

图2.5的椭圆框中，唯一没有“误差”一词的就是构建及其测量之间的不一致。心理学中的测量理论，即“心理计量”提供了涉及这类问题的大量概念。建构效度 （construct validity）就是指涉及构建内涵的计量，而“无效度”则指没有获得构建内涵的计量。例如，在“全国教育进展评估”（National Assessment of Education，NAEP）中，在测量4年级学生的数学能力时，使用的是四则运算。四则运算中的每一道题，用于测量数学能力的一些方面。这里的效度是概念性的。但如果我们知道每个学生实际的数学能力，又如何与四则运算的测量题关联起来呢？效度就是测量能够反应构建内涵的程度。

在统计术语中，效度用于个体应答者层次。它记录总体中第i 个个体的构建值（尽管不是那么容易观察到甚至根本观察不到），传统上就记作μ_i，并意指第i 个个体涉及构建的真值 （true value）。当应用某个特定的测量Y时（如用四则运算来测量数学能力），简单的心理计量测量理论注意到，其结果不是μ_i，而是别的：

$$Y_i = \mu_i + \varepsilon_i$$

即测量等于真值加上某种误差，希腊字母 ε 表示对真值的偏离。偏差正是效度概念的基础。例如在NAEP中，我们把数学能力区分为0～100分，平均分为50分。上述模型意指，在特定的数学能力测量中，第i个学生真实的数学能力为57分，但测量得到的是52分，那么测量误差就等于52-57=-5；即$Y_i=52=\mu_i+\varepsilon_i=57+(-5)$。

测试

当有人说某个受访者对调查问题的回答是测量过程的一次“测试”时，到底指什么？如果对一个受访者就同一个问题提问多次，能得到什么有价值的信息呢？答案是，“测试”是一个概念，一种应答过程的模式。这个模式假设受访者对某个问题的应答本质上就是有差异的。如果某人能抹去第一次测试测量的所有记忆，并重复同样的问题，就一定会得到有差异的回答。

对理解效度，还有一点要注意：对第i 个个体的测量可以看作同样无穷个测量之一。例如，对“在过去的6个月里有几次成为刑事犯罪的受害者”的回答，是针对所有目标受访者的测量，而不是针对一个受访者。用心理计量理论的语言来说，是无穷次测试中的一次测试。

因此，如果使用测试概念，则应答过程就变成了：

$$Y_{it}=\mu_i+\varepsilon_{it}$$

这里用到了两个下角标，i 用来表示总体中的个体序列；t 用来表示测量的次数序列。每一次测量（t）是且只是无穷可能测量中的一

次。从一次调查中获得的应答（第t 次测试获得的Y_{it} ）与真值之间就有一个只有该次测量才有的误差（ ε_{it} ）。在这个意义上，一次调查就是测量过程的一次测试t ，每次测试，从第i 个个体获得的值与真值之间的误差（要使用下角标t ，记作 ε_{it} ）都不相同。例如，就数学能力而言，使用某个测量工具，如上所述，第i 个学生有时候会得52分，但如果充分测量，也许会得59分、49分、57分；如此，相应的误差就是+2，-8，0。但我们不会多次测试，而是在概念上认为每个独立的测试一定会获得不同的结果。

根据这个简单的回答与真值之间误差的例子，我们已经可以给效度下定义了。效度 （validity）就是所有可能的测试与过程中，测量值Y_i 与真值μ_i 之间的相关性。

$$E_{it}[Y_{it} - \bar{Y}](\mu_i - \mu) / \left[\sqrt{E_{it}(Y_{it} - \bar{Y})^2}\sqrt{E_{it}(\mu_{it} - \mu)^2}\right]$$

这里μ 是针对所有测试和所有个体μ_{it} 的均值，同理，$\bar{Y}$ 是Y_{it} 的均值。最开始的E 表示期望值 （expected value）或所有测试及所有个体的均值。如果y 和μ 协变，即不管怎么变化，两者都同上同下，则测量具有高建构效度。一个构建的测量有效性就意味着与建构效度高度一致。

当我们注意到两个变量高度相关，但在单变量统计值上却不相同时，对效度的讨论会更加复杂。两个变量高度相关，但却有不同的均值。例如，如果所有受访者都对自己的体重少报5斤，如此体重的真值与报告值会高度相关，但报告体重的均值会比真值少5斤。这就是心理计量理论与调查统计误差属性的背离。

2.3.2　测量误差：理想的测量与实际的测量之间的可观察差距

图2.5中一个重要的质量要素就是测量误差（measurement error）。这里的测量误差是指测量样本获得的值与真值之间的离差。举例而言，想象一下全国药物使用与健康调查（NSDUH）问卷中的问题。“你曾经使用过哪怕是一次任何形态的可卡因制品吗？”通常的发现是（参见第5.3.5节 和第7.3.7节 ），非预期性的行为会被少报。因此，就这个问题而言，也许应答者有相关的行为（即真值为“是”），但给出的应答却是“否”，以防让他人知道后而感到尴尬。

如果这样的现象在问卷调查中是一种普遍的系统行为，那么在应答值和真值之间就会存在偏差。在上面的例子中，受访者的应答值，使用过可卡因的人数就会被低估。在统计上，我们需要引入一个新的术语用于描述应答值与真值之间的差值，如第i 个受访者的差值Y_i ，我们把对提问的应答叫作y_i 。这样，我们就可以系统描述与真值之间的偏差（$y_i - Y_i$ ）。让我们回到当前就业统计调查（CES），对某些雇主而言，在非监测性就业领域的雇工数也许为12，但对访题的应答值却为15。用调查测量理论的术语来说，就出现了偏差，即$y_i \neq Y_i$ ，就这个例子而言，$y_i - Y_i = 15 - 12 = 3$。

从测量的角度看，理论上，每个测量都是无穷个测量之一。因此，我们可以把单次测量看作一次测试。如果上面描述的应答偏差是系统性的，也就是说，如果偏差的方向在多次的测试中始终是相同

的，这就是所谓的应答偏差（response bias）。偏差（bias）是期望值（所有理论测试的值）与估计的真值之间的差值。这里有两个应答偏差的例子。在NSDUH中，就询问到的许多物质包括香烟和毒品滥用率的独立估计值而言，应答者的应答都是有偏差的，即人们总是倾向于低报自己对各种物质的滥用。部分可以解释的是，一些人担心如果报告了自己对某些物质的滥用会影响人们对自己的看法。同样，人们也发现刑事犯罪受害者的比例也是被低估的。调查值低于真值，两者之间存在系统偏差。在统计上，我们把所有调查应答值的平均值或期望值记为$E_t(y_{it})$，和前面一样，t 代表某次测试（调查）。如果出现下面的情况，也就意味着出现了应答偏差：

$$E_t(y_{it}) \neq Y_i$$

方差或变量误差

只要是变量引起的误差，在调查过程中就是可重复（测试）的。如果统计估计值与调查值不同，就是变量误差。在应答阶段，变异性会影响受访者的应答。设计、调查的合作似然率或样本特征的变异性会影响调查统计量。

不同的是，方差不能被直接观察到，因为调查不是真正的重复实验。

除了系统性低报或高报会产生有偏差的报告以外，还会有应答者不稳定的情形，因此会产生另一类误差。让我们看看消费者调查

（SOC）中的问题，“你认为现在的执业环境比一年以前更好了还是更差了？”人们对这个问题的应答，除了访题中的字词和之前问过的问题以外，应答者还会应用其他的刺激信号来应答。但搜集信号的过程，包括对之前问题的回忆，却是一个无序的过程，在每次调查中，都是不可预测的。其结果是产生针对理论测量值的偏差，通常称之为应答偏差的差异性。对这类应答误差，没有一个很贴切的概念，实际上也是低*信度*（reliability）或无信度应答的一种。调查统计学家称之为*应答方差*（response variance），用于区分应答偏差。两者区别在于，应答偏差是系统性的，会始终高估或低估；应答方差的估计值是不稳定的。

2.3.3　过程性误差：用于估计变量与受访者应答变量之间的可观察差距

在搜集到数据之后，且在进行估计之前，会产生什么误差呢？例如，分布中的明显极值也许是实际情况。全国刑事犯罪受害者调查（NCVS）的应答者可能报告说每天受到多次袭击，听起来也不大可信，如此，在某些编辑原则下，就会被当成缺损值。但如果补充信息说明受访者是某个酒吧的保安人员，那么受访者的应答就变得可信了。依据决于问题的测量结构，在编辑阶段，需要说明应该或者不应该提醒注意此类问题。是否提醒，会影响过程误差。

另一类的*过程误差*（processing error）来自于允许应答者用自己的话回答。例如在SOC中，有访题问到，“在过去的几个月里，您是否听说过执业环境向好或向坏的变化？”如果受访者对访题回答“是”，访员会追问，“您听到的是什么？”这时，访员忠实地记录

受访者应答的文本信息。例如，受访者说，“我听说我们的工厂计划裁员，我担心我是不是会丢掉工作。”类似于这样的，记录了不同受访者应答的丰富信息，但应答本身不能用于定量归纳。而定量归纳恰恰是问卷调查的主要产出。那么在我们称之为*编码*（coding）的阶段，就要把这些文本信息进行归类。例如，上述的应答也许会被归为“可能被现有工作机构解雇”。对这类测量的单变量归纳，就会把应答归入某个类别中。例如，8%的样本报告说“可能被现有的工作机构解雇”。

那么，这一步会出什么误差呢？不同的人对这些文本信息进行编码就会有不同判断和归类。由此产生的变异性（如编码方差）就是编码系统的函数。对编码人员拙劣的培训会使其始终误解文本信息，进而产生编码偏差。在统计上，如果对收入变量编码，就需要将受访者的应答偏差和编码者的编码偏差一样对待。这样，y_i 就是调查应答值，y_{ip} 就是编辑后的应答值，过程或编辑偏差就等于$y_{ip}-y_i$ 。

2.3.4　覆盖性误差：目标总体与样本框之间的非可观察差距

从图2.2的左边（测量）转向右边（代表性）的最大变化就是关注的焦点从个体应答值转向了统计量。请注意图2.5表示的是样本均值，一个简单的统计量，归纳了总体要素在个体层面的值。尽管有很多可能的调查统计量，但这里，我们用均值作为例子。

有些时候，目标总体（我们要研究的有限总体［finite population］）并没有一个与之完美匹配的方便的样本框。例如在美

国，并没有一个及时更新的居民列表可作为个体样本框。相反，瑞典有人口登记制度，对每个居民，都有及时更新的姓名与住址。在美国，常常用电话号码作为样本框，用于针对把美国居民作为研究总体的抽样。由此产生的误差来自于就具体调查的问题而言，能够通过电话联系到的美国居民的比例，以及他们在统计量与其他人的区别。低收入群体和居住在偏远农村的通常较少有家庭电话。如果调查统计量涉及有多大比例的人收到了政府派发的失业补助，那么电话调查就会导致低估。这就是统计上的覆盖性误差。

图2.6说明了两类因样本框而产生的覆盖性误差。目标总体不同于抽样框。目标总体左下部分没有出现在抽样框中，例如用电话号码作为抽样框，则没有电话号码的家户就被漏掉了。相对于研究总体而言，这种抽样框就叫*覆盖不足*（under coverage）。样本框右上角的部分不属于目标总体，但却属于抽样框总体，例如用电话号码作为抽样框，商用电话号码就可能被作为家户电话号码。这些就是不合格单位（ineligible units），有时被称之为过度覆盖（over coverage），有时被称为外部要素。

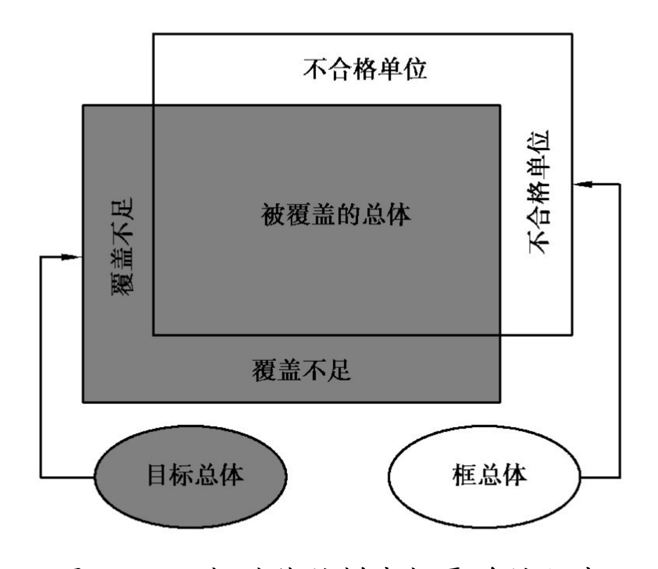


图2.6　目标总体被样本框覆盖的程度

在统计上，就样本均值而言，*覆盖性偏差*（coverage bias）可以有两个表述：目标总体没有被样本框覆盖的比例，被覆盖的和未被覆盖的总体的差值。首先，*覆盖性误差*（coverage error）是指特定统计量上的样本框与目标总体关系。在抽样之前，问题就已经存在了，不是因为要做抽样调查而产生的问题。即使我们用样本框对目标总体做普查，也存在这样的问题。因此，最简单的表述就是，覆盖性误差是指在抽样之前就已经存在的误差。如此，我们可以记述样本框均值效应如下：

$\overline{Y}$ =整个目标总体的均值。

$\overline{Y}_C$ =被样本框覆盖的总体的均值。

$\overline{Y}_u$ =未被样本框覆盖的总体的均值。

N =样本总体的总数。

C =样本框内合格的总数（被覆盖的数）。

U =所有合格的但却未被样本框纳入的总数（未被覆盖的数）。

则覆盖性偏差可以表述为

$$\overline{Y}_C - \overline{Y} = \frac{U}{N}(\overline{Y}_C - \overline{Y}_u)$$

也就是说，覆盖不足造成的误差来自于未被覆盖的比例（U/N）以及被覆盖和未被覆盖的均值差。等式左边的均值差是覆盖的均值和目标总体均值之间的差，即覆盖误差。右边则是代数，表示的是覆盖偏差

是未被样本框覆盖的目标总体比例以及被覆盖的和未被覆盖的均值差的函数。举例而言，许多针对美国家户总体的统计量，用电话框常常是不错的选择，因为没有电话的家庭比例极小，只占总人口的5%。想象一下，我们用电话调查做SOC，测量教育年限的均值，有电话的家户的均值为14.3年。对无电话的家户，即使用电话调查丢掉的那部分，受教育年限的均值为11.2年。尽管没有电话的家户的均值较低，但就覆盖性而言，其偏差是：

$$\overline{Y}_C - \overline{Y} = 0.05(14.3 - 11.2) = 0.16(\text{年})$$

换句话说，我们期待的是样本框的均值要么是14.3年，要么是14.1年。

抽样框的覆盖性误差在抽样调查中产生样本均值估计值$\overline{Y}_C$ 而不是总体均值$\overline{Y}$ ，因此抽样框的覆盖性误差属性直接影响基于样本统计量的覆盖性误差属性。

2.3.5　抽样误差：样本框与样本之间的非可观察差距

还有一个误差被引入了抽样调查统计量。由于成本或调查的不可实现性，不是样本框的所有对象都会接受测量，而是抽取其中的一些样本，作为测量的对象，其他的则被忽略了。几乎在所有情况下，这种审慎的无观察，都会在完访样本的统计量与全样本框的统计量之间形成偏差。

例如，NCVS起于全美的3 067个县，然后用人口规模、地区以及相应的犯罪行为，形成组或层。在每个层，每个县的被选概率相等，然后抽出了237个样本县或县组。所有被调查对象，都来自于这些地区。每个月从样本区抽选8 300户进行调查。

和其他调查误差一样，这里也有两类*抽样误差*（sampling error）：抽样偏差和抽样方差。如果抽样框的某些单位没有被抽选的机会或被抽选机会减少，就会出现*抽样偏差*（sampling bias）。在有抽样偏差的设计中，每一种抽选方法都会将其系统性地排斥在外。除了其调查统计量有特别的值以外，其他统计量也与框总体不同。在给定抽样方法下，如果有多个不同的抽样框组（例如NCVS中不同的县和家户），就会产生*抽样方差*（sampling variance）。每一组在统计量上都会有不同的值。

与2.3.1讨论的测量测试一样，理论上抽样是可重复的，这就是抽样方差的来源。图2.7展示了这个基本概念。图的左边表示通过不同的样本获得的不同的、可能的抽样要素，我们列出了S种不同的样本实现（realization）方式，以及每一种方式的分布状态（x轴是变量的值，y轴是该值的频次）。让我们用样本均值做例子。S个样本集中的每个样本集，都有一个不同的样本均值。表达样本方差的方式见图的右边。这是均值的*抽样分布*（sampling distribution），即样本均值的频数分布（x轴是样本均值，y轴是S个样本集中具有该均值的频数）。分布的离散性用于测量抽样方差的正态分布性。如果S个样本集的平均样本均值等于抽样框的均值，就不存在均值意义上的抽样偏差。如果分布的离散性较小，则抽样方差也较小。如果变量值为常数，则抽样方差为零。

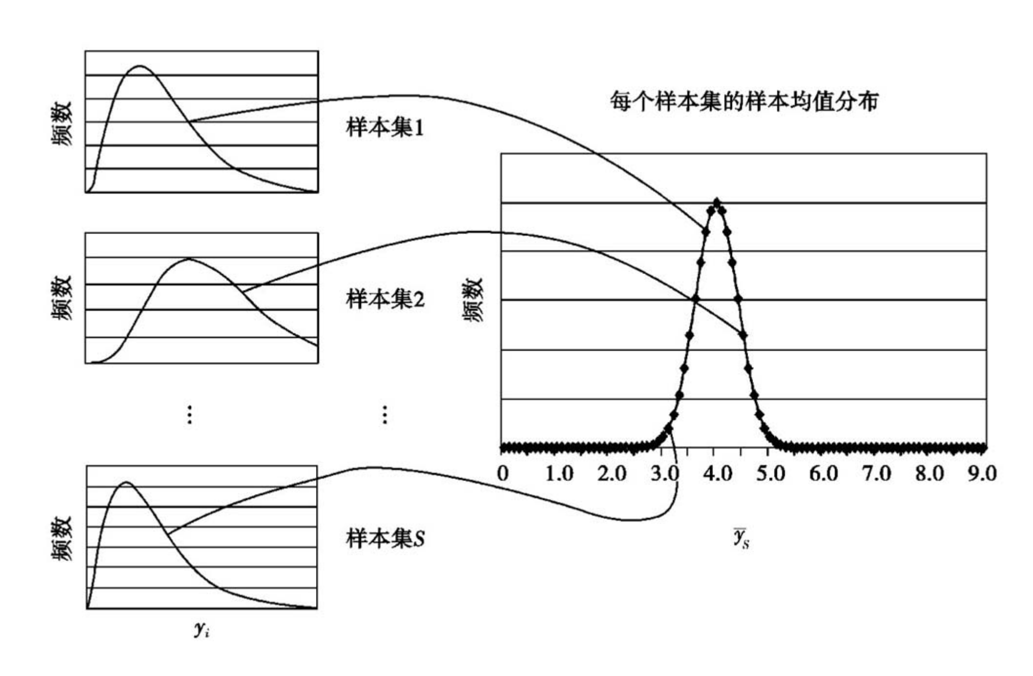


图2.7　样本集与抽样均值分布

抽样中，误差的范围是4个基本设计原则的函数：

1）抽样框整体情况是否已知，入选样本的概率为非零，即概率抽样（probability sampling）。

2）框中的关键子总体是否有样本代表，即分层（stratification）。

3）样本是在抽样框中直接、独立抽取的还是分组抽取的，即个体样本（element sample）或整群样本（cluster sample）。

4）选择多大的样本量。

如果运用这些术语，则NCVS每个月大约8 500户的样本就是分层的、整群的概率样本。抽样误差主要来自于对不同的框所设计的概率

抽样方式。

抽样误差主要来自于不同框抽选的概率性。如果每个抽样单位的被选概率是相等的，就没有所谓抽样误差问题了。在分层和非整群的条件下，抽样方差可以通过增大样本量的方式来降低。用统计术语来说就是：

y_S =抽取样本s 的均值，s =1，2，…，S

Y_C =抽样框中样本集C 的均值

这些均值可以表达为：

$$\bar{y}_S = \frac{\sum_{i=1}^{n_s} y_{si}}{n_s}, \text{以及} \bar{Y}_C = \frac{\sum_{i=1}^{C} Y_i}{C}$$

抽样方差 （sampling variance）测量的$\bar{y}_S$ 是变量对样本整体情况的反应性。常用的测量工具是样本均值与样本框均值之间的平方差，用公式表达为：

$$\frac{\sum_{S=1}^{S} (\bar{y}_S - \bar{Y}_C)^2}{S}$$

如果抽样方差高，则样本均值就很不稳定。在这种情况下，抽样误差就很高。这就意味着，任何使用这类设计的调查，有更大的可能使得从调查中获得的均值会远离抽样总体的均值。

2.3.6　无应答误差：样本与应答者群体之间的非可观察差距

除了尽力以外，调查中，不是所有的应答者都可以应答的。有时候，对非生命体要求100%的应答率（例如人们的医疗记录，住宅）。但对生命体的调查而言，从来没有出现过100%的应答率。例如在消费者调查（SOC）中，每个月都有30%～35%的样本回避或拒绝接受访问。在全国教育进展评估调查（NAEP）中，大约17%的学校拒绝参与，在参与的学校中也有11%的学生没有被访问到，要么是因为缺席，要么是因为家长拒绝。在奇数年，根据法律，要求学校4年级和8年级的学生参与阅读和数学测试。2007年，100%的学校都参与。

无应答误差（nonresponse error），就是依据实际应答数据和依据假设的完全应答数据计算的统计量之间的差值。例如在NAEP中，缺席学生的数学或组词能力较差，NAEP就会出现无应答误差，即获得的成绩系统性地被高估。如果无应答率很高，那么高估的情形就会很严重。

大多数针对调查参与性的研究者关注的都是无应答误差。与此相关的统计量已经在2.3.1中说明：

$\bar{y}_s$ =全部选定样本的均值。

$\bar{y}_r$ =在第s 个样本集中，应答的均值。

$\bar{y}_m$ =在第s 个样本集中，无应答的均值。

n_s =第s 个样本集中的样本数。

r_s =第s 个样本集中应答样本的数。

m_s =第s 个样本集中无应答样本的数。

那么，针对全部样本的无应答偏差就是

$$\bar{y}_r - \bar{y}_s = \frac{m_s}{n_s}(\bar{y}_r - \bar{y}_m)$$

因此，样本均值的*无应答偏差* （nonresponse bias）等于无应答率（无应答样本占总样本的比例）与应答均值减去无应答均值之差值的乘积。这说明仅仅应答率，并不是质量指标。高应答率的同时可能也有高的无应答偏差（如果无应答者对调查变量非常显著）。因此，高应答率只是减少了无应答偏差的风险。

2.3.7　调整偏差

图2.5非观察误差部分的最后一步是事后调整。人们做了种种努力来改善样本估计准确性，包括覆盖面、抽样以及无应答误差。在某种

程度上，他们与2.3.3讨论的涉及每个应答值的编辑工作具有相同的功效。

调整会使用到目标总体或框总体，以及应答率的信息。调整会给在最终数据集中代表性不足的样本更大的权重。想象一下，你对美国的个人犯罪率有兴趣，全国刑事犯罪受害者调查（NCVS）在城镇地区的应答率为85%（即85%的受访者回答了问题），其他地区的应答率为96%。这就意味着，在应答者中，城镇的代表性不足。对此，一种办法就是建立两个权数，即w_i=1/0.85给城镇应答者；w_i=1/0.96给其他地区的应答者。调整后的样本均值为

$$\bar{y}_{rw}=\frac{\sum_{i=1}^{r}w_i y_{si}}{\sum_{i=1}^{r}w_i}$$

即在计算城镇应答者的均值时给更大的权重。由调整均值所带来的误差则与目标总体均值有关。

$$(\bar{y}_{rw}-\bar{Y})$$

误差会与调查的样本和数据的应用有关。也就是说，调整一般也会影响到估计值和方差的偏差，且随样本和对数据应用不同而不同。最后，事后调整会减少覆盖性、抽样以及无应答误差，同时，也可能增加误差。对此，第10章有详细讨论。

2.4　从全局着眼

本章是从调查的三个视角开始的。第一个视角如图2.2，说明了设计的几个阶段，从抽象的概念到具体的行动步骤。第二个视角如图2.4，展现了调查的步骤，从开始到结束。第三个视角如图2.5，说明了调查的质量特征，涉及实地工作的每个步骤，以及表述质量的不同概念。质量讨论涉及集中关注了调查两个估计值：观察误差和非观察误差。观察误差涉及建构、测量、应答以及调整后的应答之间的差距。非观察误差涉及目标总体、抽样框、样本、应答者之间的差距。

这些误差有些是系统性的，即在调查中系统性产生的影响（如无应答），我们把这些叫做偏差（biases）。另一些则是在涉及变量值时随机出现的误差（如效度），我们称之为方差（variance）。实际上，正如本书后面的章节中展现的，所有这些误差来源都是系统性的和随变量发生的，因此也包含了偏差和方差。

现在读者应该理解了定量调查也有量化的质量特征。让我们再看看图2.5，说明了一个简单的样本均值相关的术语随调查步骤的变化。本书的其他章节还会继续使用这些术语。大写字母代表总体特征。如果不存在针对特定目标总体问题，大写字母也用于讨论测量。在讨论通过样本推论到目标总体时，大写字母用于标示总体要素，小写字母用于标示样本要素。变量的下角标用于标示子总体的所在（如i 标示第i 个人，或标示调整后的状态，如w 标示加权）。

2.5　各种统计维度上的误差

上述讨论仅仅分析了一种调查统计量——样本均值，以说明误差原理。调查中还有许多其他的统计量（如相关系数、回归系数）。

对调查的两种应用及其相关的统计量，在这里值得一提。

1）描述性应用，即属性在总体的分布状态，群体在总体的规模，或某个测量的均值。

2）分析性应用，即什么导致了现象的出现或者两个属性之间的关联度有多大。

许多调查的目的在于了解分布信息，如总体的特征、观点、经验、想法等，且通常会报告均值。例如NCVS可能报告说，在过去的1年，5%的人报告过汽车被盗。

相反，类似于“女性较男性更常去看医生”“共和党人较民主党人更喜欢去投票”，或者“与65岁以上的人比较，年轻女性更容易成为刑事犯罪的受害者”之类的陈述，都是关系性的。在某些情况下，描述关系的程度是非常重要的。研究者也许会说，家庭收入与投票可能性之间的相关性为0.23。另一种可能是，随着家庭收入的增加，对教育的投入随之增加，这就是所谓的因收入产生的“模型”：

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

式中，y_i 是第i 个人的收入；x_i 是第i 个人的受教育年数。

模型中要检验的假设是早期教育获得的收入回报要大于后期的投入。如果相关系数为β_1是正向的，β_2是负向的，就意味着支持原假设。这是运用调查数据进行因果分析的例子。在这种情况下，例子关注的就是教育获得是否导致收入获得。

类似于相关系数和回归系数都涉及上面讨论过的同一个调查误差吗？是的。如果用调查数据去估计统计量，就都会涉及覆盖面、抽样、无应答、测量误差，就像简单样本均值一样，尽管数学表达式不同。对大多数测量关系的统计量而言，统计误差是两个变量之间关系的特质（如协方差和方差的特质）。

在分析性统计和计量经济学文献中，理解与因果推论相关的统计误差是很重要的。在这些领域，表述误差的语言与调查统计中有所不同，但概念含义却是相似的。

因此，所谓的分析，就是把要回答的问题与其相关的各种误差纳入考虑。偏差，无论来自应答者样本，还是来自应答本身，在描述分布时都是首要考虑的问题。如果调查的主要目的是估计特定犯罪受害者所占的比例，且该类犯罪被系统地低报了，如此就会直接影响调查目标的达成。相反，如果关注的仅仅是年轻人或者年老人更容易成为受害者，即使有些稀少犯罪被低报，也不会对调查目标的达成构成重要影响。如果所有年龄组的偏差一致，即使所有受害估计值都偏低，研究者也可以获得年龄与受害之间关系的有效结论。

2.6　调查质量的非统计评价

除了上述总调查误差视角以外，还有三个涉及调查估计质量的评价视角。这三个视角都涉及对调查的应用拟合度（fitness for use）的最大化期望。应用拟合度是指，同一个调查对不同的使用者而言，具有不同的信息含义。对某个使用者而言，高准确度的调查估计值是好决策的信息基础；对另一个使用者而言，与总体排序一致的大致调查估计值也许就足够了。这就意味着，对第二个使用者而言的“好的”调查对第一个使用者而言不是“好的”。高应用拟合度意味着为特定应用提供了其所需要的信息。

第一个视角是“可信度”（credibility），即数据使用者会从各个角度来评价调查提供的信息，而测量的切入点会影响到已知调查方向的产出。中央政府统计机构试图获得中性的、客观的数据。科学家则运用调查方法（1）记录设计和执行的每个步骤，以便可以重现结果，然后（2）说明影响到其结论的弱项。这两个方面都是试图提升调查数据的可信度。

第二个视角是“关联性”（relevance）。如果调查测量的构建与数据使用者主要关注的焦点非常相似，调查与数据使用者就具有关联性。有时候，两者之间是有差距的。例如，数据使用者也许希望分析经济困难的普遍性，以及在个体层次因为经济困难带来的情绪和身体的不适。而由于政府救助机制的影响，人们也许不因为失业而感受到不适，数据使用者也因此会在关联性方面提出批评。这与建构效度与关联性之间的关系类似，关联性关注不同的构建，而建构效度则与构建及其测量有关。

第三个视角时“及时性”（timeliness）。调查是否符合数据使用者的需求，取决于数据使用者在需要信息时是否有相关的信息可用。例如，2009年3月消费者调查（SOC）描述的消费者信心指数对

2010年3月的宏观经济就没有任何意义。调查的及时性完全取决于使用者。

的确，所有三个视角只有在数据使用者基于特定的目的使用数据时才可以给出评价，也是调查总体误差之外的视角，不受实际调查方法的影响（后面还会进一步讨论）。无论如何，在考虑到同一个数据面对不同的使用者时，这些都是重要的维度。

2.7　小结

抽样调查有赖于两类推论，从问题到构建，从样本统计到总体统计。推论包含两个相互关联的步骤集：获得结构性访题的答案，识别和测量从目标总体中获得样本单位。

除此以外，每个步骤都不会完美，并因此在调查统计中产生误差。误差包括效度问题，即测量与构建之间的差距。使用测量工具时产生的误差被称为“测量误差”。在数据清理和准备时也会产生误差，即所谓调整误差和过程误差。在使用样本框推及目标总体时还会产生覆盖性误差。如果样本只是框总体的一个子集，则会产生抽样误差。如果未完成应做的测量，就会产生无应答误差。在使用已获得的数据来描述整个目标总体时，还会产生事后调整误差。即使使用同一个调查数据，但对不同的统计分析而言，上述的各类误差也会有不同的影响。

本章给读者介绍的是实地调查方法的基本要素，包括简要的解释和概念。后面我们会进一步给出相关的研究文献以及在测量人的时候和调查大总体特征时的新原理，也会说明建构的理论如何影响了日常

的调查任务和执行。调查实践的目的，就在于改善调查统计的质量（或降低调查的成本）。通常，调查实践也会通过筛选提供新的测量。

关键词

偏差（bias）

整群样本（cluster sample）

编码（coding）

构建（construct）

建构效度（construct validity）

覆盖性偏差（coverage bias）

覆盖性误差（coverage error）

可信度（credibility）

个体样本（element sample）

误差（error）

非观察误差（errors of nonobservation）

观察误差（errors of observation）

期望值（expected value）

有限总体（finite population）

应用拟合度（fitness for use）

补值（imputation）

不合格单位（ineligible units）

推论（inferenc ' e）

选项缺损值（item missing data）

测量（measurement）

测量误差（measurement error）

数据搜集模式（mode of data collection）

无应答值（nonrespondents）

无应答偏差（nonresponse bias）

无应答误差（nonresponse error）

观察单位（observation unit）

奇异值检测（outlier detection）

过度覆盖（overcoverage）

事后调整（postsurvey adjustment）

概率抽样（probability sampling）

过程性误差（processing erro）

实现（realization）

关联性（relevance）

信度（reliability）

应答值（respondents）

应答（response）

应答偏差（response bias）

应答方差（response variance）

抽样误差（sampling error）

抽样偏差（sampling bias）

抽样框（sampling frame）

抽样方差（sampling variance）

统计值（statistic）

分层（stratification）

目标总体（target population）

及时性（timeliness）

总调查误差（total survey error）

真值（true values）

覆盖不足（undercoverage）

样本无应答（unit response）

效度（validity）

加权（weighting）

进一步阅读资料

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality* , New York: Wiley.

Groves, R. M. (1989), *Survey Errors and Survey Costs* , New York: Wiley.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys* , New York: Wiley.

Weisberg, H. (2005), *The Total Survey Error Approach : A Guide to the New Science of Survey Research* , Chicago: University of Chicago Press.

作业

---

1. 最近的一篇报纸文章报道说，“上个季度，手持电子设备（如黑莓、PDA等）的销售增长了10%，而笔记本电脑和台式机的销售则没有变化。”这篇报告的依据是一项网上调查，在126 000个应答者中，9.8%的人说“在今年的1—4月购买了手持设备”。

    调查中，把从美国的5大商业性互联网接入商获得的电子邮件地址作为抽样框，通过发送电子邮件请求的方式，进行调查。数据搜集的时间从2002年的5月1日开始，历时6周。总应答率为13%。

    假设作者要推论到18岁及以上的可能购买行为。

    （a）目标总体是什么？抽样框中的总体是什么？

    （b）根据本章的以及你阅读的文献，简要讨论，调查设计如何影响以下的误差来源：

    •覆盖性误差

    •无应答误差

    •测量误差

    （c）在不改变调查时长或调查模式（即电脑辅助或自访）的前提下，如何能降低上述误差？对每个误差来源而言，给出一项可以降低误差的建议。注意，要依据你的阅读以及课堂资料。

（d）如果在未来要降低调查成本，研究者试图将样本量减半，即仅使用两大互联网接入商的电子邮件地址。如果这样做对抽样误差和覆盖性误差有什么影响的话，影响是什么？

2. 说明调查估计中覆盖性误差和抽样误差的差别。

3. 假设你已经阅读了关于覆盖性误差、无应答误差以及测量误差的文献，请设计一个抽样调查，并试着说明降低一项误差会导致另一个误差增加的情形。在建构样本后，做一个方法研究的设计，看看对一项误差的降低是否真的导致了另一项误差的增加。

4. 本章讨论了观察误差和非观察误差。

（a）列举3项影响从样本推论到目标总体的误差来源。

（b）列举3项影响从应答值推论到构建的误差来源。

（c）针对列举的每个误差来源，说明其是否潜在地影响了估计方差、估计偏差，或两者。

5. 对下述每个设计决策而言，识别误差来源会有重要影响。每项决策至少会影响到2项误差来源。请简要（2～4句话）回答每个问题。

（a）为了解美国人身体残障的状况，在抽样框中是否包括机构性人员（如住院病人、正在监狱服刑的囚犯、居住在军营的军人）。

（b）为了解接受社保福利的老年人住房情况，是否采用邮寄式自访问卷。

（c）为了解对居家消费品的满意度，是否多次打电话给受访者劝其参与调查。

（d）为降低调查成本，是否使用现有工作人员调查健康维护组织（HMO）的病人，并增加样本规模。调查的主题是医疗服务的满意度。

（e）是否将原定的在1月1日—5月1日完成的、对少儿家长使用育儿设施的调查延长到1月1日—8月1日。

（f）在消费者支出调查中，是否把囚犯和住院病人纳入抽样框。

（g）在调查良性平等的态度时，是否使用现有的女性访员而不是雇佣新访员。

（h）在毒品使用调查中，是否把面访改变为邮件访问。

6. 对下面的每个问题，请简要说明设计者试图测量的构建。在一些情况下，会有多个测量目标。

（a）你多大？

（b）你结婚了吗？

（c）你有车吗？

（d）你的收入是多少？

（e）上次总统选举时，你投票了吗？

（f）你认为自己是民主党、共和党，还是政治独立人士？

（g）在未来的12个月里，你认为经济状况会更好吗？更差吗？还是不会改变？

（h）你认为自己是一个快乐的人吗？

（i）医生是否告诉过你有高血压？

（j）你如何评价医生的诊治能力：非常好，好，一般？

（k）上周，你自己做过饭吗？

7. 从推论的视角出发，对样本无应答需要关注的核心问题是什么？

3　目标总体、抽样框以及覆盖性误差

3.1　导言

抽样调查用于对明确界定的总体进行描述或推论。本章从理论和实践两个方面讨论对总体的界定和识别。总体的基本构成单位是“要素”（elements）[(1)]。所有要素的集合，就是总体。在大多数家户总体中，要素就是居住在家户中的个人。在全国教育进展评估（NAEP）调查和其他学校样本调查中，要素就是学校总体中的学生。在类似于当前就业统计调查（CES）这样的商业调查中，要素是机构。要素可以有多种不同类型的单位，即使在同一个调查中也如此。例如在家户调查中，除了个人以外，也可以在个人居住的家户层面做推论，甚至在家户所在小区层次做推论，或者家户加入的教堂即教区层次做推论。简言之，对不同总体的统计描述，如果其测量单位之间互有关联的话，可以来自于同一个调查。

在一般的研究工具中，调查的独特性在于其清晰界定的总体。例如，如果要做生物医药实验，研究者更关注的是实验引物和条件，而不是对研究总体的识别。这类研究中隐含的假设是实验的主要目的是找到运用引物获得假设效果的条件。第二位的才是所谓的变异性问题。由于把调查作为工具的目的在于描述固定的、有限的总体，因此调查针对的就是特定的、用于研究的总体。

3.2 总体和框

“*目标总体* ”（target population）是调查研究者试图通过样本的统计分析进行推论的要素组。目标总体的规模是有限的（即至少在理论上是可数的）。目标总体也具有某种时限性（即会存在于某个特定时限内）。目标总体还是可观察的（即是可接触的）。对目标总体特征的这些期待，目的在于对调查统计的含义获得清晰的理解以及使得调查本身可重复。

目标总体的定义中，一定要包括总体中的某种要素单位和时间段。例如，许多美国家户调查的对象是在美国境内家户中居住的年满18岁或以上的成年人。一个*家户* （household）实际包括了所有居住在居住单元中的个人。*居住单元* （housing unit）可以是一幢房子、一个公寓单元、一个流动住所、一个房间组、单个房间等已经被占据或即将被占据的独立生活空间。分离的居住空间是指建筑物中虽然人们从同一个空间进入建筑物但却各自生活起居的空间。占据生活空间的人们可以是一个家庭、独自一人、两个或者多个家庭，或者任何其他与个人相关或不相关的共同生活形式。在美国，在某个时点，并不是所有的个人都是成年人，也不是所有的成年人都住在居住单元中（一些人居住在监狱、长期提供护理的场所，或军营）。

在美国，也不是所有全国性家户调查都选择这类目标总体。有些调查仅仅涉及美国大陆上的24个州以及华盛顿特区。另一些调查则可能包括军人或军事基地，还有一些调查会限定在美国公民或母语为英语的人口。

由于总体总是随时间变动的，因此目标总体与调查时间也密切相关。入户调查常常要持续几天、几周甚至几个月。在家户层面，人的进出几乎每天都发生，因此许多入户调查的目标总体就是调查期间家户人口的一个集。在实践中，不少调查的家户成员在第一次联系时就已经确定了。

在搜集数据的实际操作中，常常会遇到一些禁区，进而也影响到调查总体的确定。例如，在一些国家，由于局势不稳，就会使得在一些地区搜集数据变得几乎不可能。这些区域的范围也许很小，在抽样时，就被从总体中排除出去了。排除受限制总体后的总体，有时叫作“调查总体 ”（survey population），并不就是目标总体，而是被用来实际搜集数据的总体。例如，CES的目标总体是某个月所有雇用了人员的机构。但是，调查总体却是开工了几个月（即在一段时间内存在）的雇主。调查机构也许会在技术性文档中注明目标总体（例如存活着的个体）与调查总体（除动荡地区以外的存活者的个体）之间的差距。

资料集或“抽样框 ”（sampling frame）就是用来识别目标总体中的要素的。抽样框是一个列表或用于识别一个目标总体中所有要素的过程。抽样框也许是一个区域的地图（包含了要素）；事件发生的时间段，或向内阁提交文件的记录表，或者其他什么的。最简单的抽样框就是总体要素的一个列表。有些列表是现成的，如专业性组织的成员，具体城市或县域的商业机构、医院、学校，或其他机构。一些国家的个人登记或地址登记，也可以用作抽样框。

尽管有很多的总体，但不是所有的列表要素都是可用的。例如在美国，很少能找到一个州可以给出所有在校学生、在狱犯人的列表，甚至没有一个县的成人列表。也许有一些具体机构成员的列表或要素

的集，但却很少有跨机构的成员列表或综合性的列表。有时候，还必须在数据搜集期间制作列表。例如，对住户调查而言，住户列表就常常是不可得的。在区域抽样中，在抽样的一个或多个阶段，常常精确定义一个小的区域例如城市的街区，而不是更大的区域，然后，就可以让人到选中的街区去制作所有住户的列表。

如果可用的抽样框部分或全部不在目标总体中，研究者就会遇到下列问题：

1）重新定义目标总体，以更好地匹配抽样框。

2）在对原目标样本的统计性描述中，允许可能存在的覆盖性误差。

在家户电话调查中，用于抽样的框是电话号码集，经常能见到重新定义目标总体的现象。即使理想的目标总体是居住在美国的成年人，使用电话调查的方式也许会促使研究者将目标总体改变为有家户电话的成年人。此外，研究者也可以保持原目标总体，但需要在文档中说明，在美国约有2%的成年人家户没有被覆盖，因为他们没有电话。采用新目标总体的动力主要来自于研究者感兴趣的群体不在总体中。坚持使用全部家户作为目标总体就意味着要面对针对覆盖性的批评。当然，如果目标总体是全部家户，这也是电话调查与其他调查相比的弱势和不完善的地方。

有意思的是，上面讨论的观点还会影响到NAEP，即对美国在校学生的调查。如果目标总体是美国所有的在校学生，而抽样框仅仅是美国公立学校，情形会如何？由于就读私立学校的学生大多来自富裕家

庭，他们的语数成绩常常要高于公立学校。如果要调整目标总体以匹配抽样框（即仅仅涉及公立学校），就会招致这样的批评，即没有覆盖所有学生，特别是，不是美国的政策制定者所感兴趣的。如果使用全部学生（涉及所有学校）作为目标总体，由于没有调查私立学校，就会导致覆盖性误差。总之，前一个问题涉及用户的需求差异，后一个问题涉及调查操作的统计弱项。

3.3　样本框的覆盖性

尽管目标总体和调查总体之间是可以有差异的，但统计关注的中心问题是，抽样框（可用的抽样选择）在多大程度上覆盖了目标总体。第2章的图2.6说明，抽样框与目标总体的匹配产生了三个潜在的问题：覆盖性，覆盖不足，以及不合格单位。

如果目标总体的要素在抽样框中，也就意味着目标总体被*覆盖* 了（covered）。当然也有目标总体没有或不可能被抽样框覆盖的情形，这就叫*覆盖不足* （undercovered），即总体中的合格成员没有出现在任何用于调查的样本中。另一种情形是*不合格单位* （ineligible units），即抽样框中出现了非目标总体的单位（例如，在家户电话调查目标总体中出现了商业单位的电话）。

如果抽样框与目标总体要素能够一一对应，那就是完美的抽样框。在实践中，完美的抽样框是不存在的，总有各种情形会破坏一对一的完美匹配。

对一个抽样框而言，考察四个方面的问题是重要的。其中，两个方面已经在上面做了简要的讨论，即覆盖不足和不合格单位问题。另

外两种情形的一种是，抽样框中的单位可以匹配到目标总体，但匹配的却不是唯一的。重复 （duplication）指的是多个样本框单位指向目标总体的单一要素。在抽样调查中，如果存在重复问题，就会出现对目标总体要素过渡代表的问题。“汇聚 ”（clustering）指的是目标总体中的多个要素与抽样框的单一要素关联。在抽样调查中，样本规模如果遇到汇聚的情形，就会过小或过大。当然，也存在上述两种情形的交叉，即多个抽样框要素与多个目标总体要素关联的情形。果真如此，情况就更复杂了。这里，只是简单地关注重复和汇聚。

3.3.1　覆盖不足

在抽样调查中，对覆盖不足最大的担心就是覆盖误差。在任何抽样调查中，由于撇下了部分目标总体，进而就会出现非观察误差。例如，在家户电话调查中，目标总体是所有家户中的个体，如果采用电话调查，由于没有任何一个电话抽样框会纳入没有电话的家户，进而出现了覆盖不足的问题。在行为风险因素监测系统（BRFSS）和消费者调查（SOC）中，就有这种情形。在电话抽样调查中，电话模式即从固定电话到移动电话的转变，也会造成覆盖性问题。在世界上的许多国家，由于使用电话会持续产生费用，穷人就会被漏掉。在一些国家，移动电话在逐步取代固定电话，如果抽样框限于固定电话，年轻人就会被漏掉，因为他们更多地使用移动电话。正如我们将在3.6节中要讨论的，对未覆盖的统计计算（以普查还是以抽样调查为基准），取决于抽样框之间的比较。

导致覆盖性问题的因素与抽样框的建构有关。建构抽样框的过程也许是调查设计可以控制的，也许超出了控制的范围（如抽样框来自

于外部资源）。例如，在一些家户调查中，调查样本最初来自于区域（如县、街区、列举区域，或其他地理单位）列表，然后才是选中区域内的住宅单元列表，最后才是住户内的个体列表，这些样本被称为区域框样本（area frame samples）或区域概率样本（area probability sam-ples）。抽样的三个阶段都可能出现覆盖性问题。

在区域概率抽样设计中，每一个选中的区域都会被用作下一阶段的框，由抽样人员用居住单位（通常运用地址进行识别）来建构住宅抽样框。抽样人员在样本区域如街区或街区组，按照指定的方法将所有的居住单位做成一个列表。制作住宅地址列表看起来容易做起来难。对于运用如街道、马路、铁道、河流、其他水体等识别性参照物的区域，某个居住单位是否应该算作区域内的或是否应该进入列表是相对容易的问题。对于想象中的分界线如山顶或山崖等自然参照物，则常常会有争议，比较难确定；某个住宅在或不在区域内，也较难确定。如果边界确定错误，分散的居住单位就很容易落掉。这也属于未被覆盖的总体。

在所有情况下，识别居住单位并不是容易的事情。一个居住单位，作为住人的区域，应该有独立的进出通道，在结构上也应该有独立的餐饮空间。独栋住宅或居住单位比较容易识别。但对搭建的居住单位，如果有障碍物遮挡视线，就不大容易区分了。有安保设施的小区或建筑物，也很难识别。无论荒野的乡村还是拥挤的城镇，如果居住单位不在视线内，也很容易被忽略。在制作列表中，在入口处不是很容易看到的窄条排列的居住单位，也很容易被忽略。不管哪种类型的被忽略，都会造成对目标总体的覆盖不足。

多居住单位建筑结构中居住单位识别也很困难。对于特定的建筑结构而言，从外部是很难看清其具体居住单位的。此时，就常常会用

到可见的邮箱、水表、电表、煤气表、入口等，这些都是居住单位的线索。但在这些情境下，那些隐藏着的入口也许就被忽略了。

有些情况下，还需要具体的规则来确认某种居住安排是否为居住单位。例如一些文化中不常见的集体性居住安排。在居住安排中，也许只有一个出入口，一个共用的厨房间，但却有依出生或收养或婚姻安排的卧室。这时要考虑的是，这些人是否属于家户，按照居住安排列表，还是按照每个卧室列表？

还要识别某些机构，建立列表规则。有些机构容易列表，如监狱或医院。抚养机构也是要识别的对象，尽管机构本身被排除在外，但机构内的居住单位却需要识别。另一些机构却不大容易识别。例如在监狱系统中有暂时性的羁押场所，嫌疑人可能处于被羁押状态而不是在监狱处于被监禁状态。是否要把羁押场所列入居住单位列表，需要在程序上说明。同样，医院或医疗场所也有附属的康复场所为需要进行康复的人提供康复或护理服务。如果制作抽样框的工作人员不确定是不是应该将其纳入列表，那么就有可能被落下，由此也会在调查统计上产生覆盖性问题。

在家户调查中，另一种容易产生的覆盖不足问题就是，抽样框通过地址或电话提供的是居住单位，而不是居住单位中个体的识别信息。在有户籍登记的国家，常常采用的是户籍抽样框，而不是居住单位抽样框。在人口普查或运用地址框但目标是个体的抽样中，还要做一个涉及个体的小抽样框。访员要做一个居住单位内居住成员的列表，如果列表没有准确地反映居住单位内的个体，就会产生覆盖性问题。

家户调查中的个体框所列出的一般是家户中的居住人口。对居住人口的界定一定要清楚，如此，访员才可以根据可获得信息判断是否应该将某个人放在个体框列表中。实践中，有两条基本准则可以用来确定是否为居住人口。在人口普查和一些调查中使用的事实认定准则（*de facto* residence rule），即调查前一晚，在居住单位中过夜睡觉的人，都算。这一准则主要用在短期数据搜集中，对那些经常更换住处的个体而言，这种方法也避免了覆盖过度。在短期内，一个人如果不止在一个居住单位居住，那么在样本中就可能被过度代表。事实认定准则比较容易实施，因为界定比较清楚。尽管某个人平时都在家住，但前一天却居住在某机构（如宾馆），就会产生覆盖不足。

调查中的另一个准则就是法律认定准则 （*de jure* residence rule），即认定居住在居住单位的“常住人口”。这个准则可以应用于很多情形，但有些情形下的使用，也比较困难。有些人的工作性质决定其出差较多，如销售代表、卡车司机或飞行员，这些人算不算常住人口 （usual residence），就不清楚了。如果人家告诉你说，如果不出差就住在那儿，大多数时候（如前一年或前一个月），就可以运用法律认定准则。即使有人准备将某居住单位作为其住所（如刚刚迁入），根据法律认定准则，这些人也是常住人口。

Mulry（2007）论美国人口普查的覆盖性

Mulry对美国2000年人口普查覆盖性的各方面进行了总结。

研究设计：采用严格质量控制的大样本调查并与普查记录进行匹配，用于测量家户和个体是否被纳入了人口普查。此外，人口分

析部分还采用行政性资料如人口变动信息：出生、死亡、迁入、迁出。最后，还采用计算机匹配普查中的重复记录。

研究发现：与对前次人口普查中未被纳入的一般性发现不同，此次研究发现了被重复计算的净覆盖过度，数量为总人口的0.5%。许多的重复发生在邮寄返回问卷与面访问卷之间，原因来自于住户搬迁、度假、学生流动以及共同照看儿童等。而覆盖不足的主要是租房人口，说非西班牙语的非洲裔人口以及男性人口（年龄在18—49岁。）

研究的局限性：对净覆盖过度人口的估计仅限于居住在居住单位的人，不包括居住在集体宿舍的人。人口分析方法的结果显示出少量的覆盖不足，与调查方法的结果比较，差异不大。

研究的影响：这项研究说明了运用新技术对测量重复计算人口方法的改进，以帮助我们理解多居所家户如何影响了家户人口计数以及如何将其定位于一个居所位置。

美国的普查和一些调查，都采用这样的方法，其覆盖性如何，都有详细的记载。年轻的男性（18—29岁），特别是少数族裔群体，与家庭的关系比较松散。一周里，他们可能有几天和父母住在一起，有几天和朋友住在一起。同样，比较贫穷家户里的孩子也是这样。尤其是双亲不全的家庭，孩子们有时候和母亲住，有时候和祖父母住，有时候和父亲住，有时候和其他亲戚住。在这样的居住单位中，如果访员询问，“谁住在这儿？”，就有可能不等比例地忽略掉了这些人（参见Robinson，Ahmed，das Gupta，and Woodrow，1993），并因此造成覆盖不足。

有时候，居住在同一居住单位的人并不是国家法律认可的。例如，根据法律，租住协议规定一个居住单位只能一个家庭5口人住，但穷人们为了分担费用也许好几家住在一个居住单位，而且不愿意说明有其他家户与其住在一起。还有，如果福利政策排除了已婚配偶，那么未婚的女性就不会告知居住单位中有男性居住者。这会系统地丢失某一类人口（de la Puente，1993）。

在某些文化中，某些人不算成常住人口，尽管按照法律认定准则他们就是常住人口。例如，婴儿就不被算作常住人口。的确，在调查方法中，传统的家户界定和总体之间的匹配，就是人们共识的部分。关键是把家户用作对个体层次进行抽样调查的单位。如果有人只调查居住单位，在操作中就得认真对待了。

在机构调查中，机构的新建、合并、注销是影响覆盖不足的重要因素。机构的定义，特别是针对特大或特小的机构，常常很难。分处多地的企业，例如连锁店，也许应该根据地理位置，将其当作多个企业。有多个办公地点或工厂、仓库、运输点，也需要单列。调查认定的机构和机构自己的认定也许很难一致。

有些机构存在的时间很短，或太小以至于没有被列在抽样框中。例如当前就业统计调查（CES）就会漏掉新建立的用人单位，也许是几个月。有时候，机构框会基于行政注册登记信息。不过，行政性信息也许更新不及时或不完全，尤其是对新建机构而言。机构的合并和拆分使得行政性记录变得复杂，并会导致覆盖过度和覆盖不足并存。

覆盖不足是难于识别和解决的问题。如果抽样框中没有出现总体要素，可能就要使用额外的框来识别（参见3.7.3节 的多框调查）。在电话家户调查中，如果使用区域抽样框，就可以覆盖区域内的所有

家户，即使是没有电话的家户。技术上，还可以通过受访者报告其他总体要素的方式来提高抽样框的覆盖性（参见3.7.2节 的多重性技术），但在美国的家户调查中，由于涉及隐私问题，通过受访者报告来获得他人的信息变得越来越困难了。

3.3.2　不合格单位

有时候，抽样框所包含的要素并不是目标总体的一部分。例如，在电话号码框中，不少号码并不在使用或不是家户电话，如此，使得对家户目标总体的运用变得异常复杂。在区域概率调查中，地图资料中有时候包括了目标区域之外的单位。当调查人员访问样本区域并做居住单位列表时，可能会把空宅或非住宅纳入其中。

当访员建构住户单位内的家户成员框时，常会使用常住人口定义，而常住人口定义不一定与提供信息的人所定义的家户一致。父母常常会认为不在家居住的学生也是家户成员。但在很多调查中，常将其纳入学校框中。提供信息的人还常常会将租住其一间房的租客排除在住户单位人口之外。研究表明，由离婚父母共同照看的孩子常会被重复计算，即同时出现在父亲和母亲的居住单位中。

尽管覆盖不足是难题，但如果问题不是太复杂，抽样框中的“不合格单位 ”（ineligible unit）或“外在单位 ”（foreign unit）便不是难解决的问题。如果在抽样前就发现了外部单位，处理它几乎不需要成本。更常见的是，直到正式调查时才发现外部的或不合格的单位。如果数量不多，那么在抽样后，通过识别可以将其从样本中剔除。如果直到存在大量的外部单位，即使是大概，只要是事前知道，

就能选择额外的单位，并预估要剔除的单位。例如，事前就知道电话号码簿中大约15%的家户电话已经不再使用，如果要抽100个家户电话号码，抽样时就需要抽取100/（1-0.15）=118个电话号码，即预估其中有18个号码已经不再使用了。

如果外部单位的量很大，抽样框是不是值得继续使用，就需要认真考虑了。例如，在美国的电话家户调查中，如果一个框中包括了所有地区号码组合（在美国10位电话号码中的前6位）。基于抽样框的调查通常被称为“随机数字拨号调查 ”（random digital dialed surveys）。如果所有可能的10位号码中，有85%的已经不再使用了（外部单位），就不值得继续使用这个框了。因为，清除外部单位是一件耗时的工程，也许其他的抽样框或抽样技术比之更加有效率（部分方法参见4.8节 ）。

3.3.3　目标总体要素在抽样框中的汇聚

正如之前所提到的，框对总体（汇聚）或总体对框（重复）的多重对应 （multiple mappings）是样本选择中的问题。运用电话号码簿（抽样框）获得居住在有电话的家户（目标总体）内的成人样本正说明了这类问题。

电话号码簿中的列表依据的是居住者的姓、名、地址。如果要从这个框中抽取成人，马上会遇到的问题是如何获得合格的个体。汇聚（clustering）是指同一个框要素代表了多重目标总体要素。电话簿中的一个电话可能代表了一个、两个或三个成人。

图3.1展示了汇聚的情形。图的左边表示7个不同的目标总体要素（人），家里有电话的人。Smith一家的电话号码为744-555-1000，电话号码是抽样用的要素。Smith家的所有人只有一个电话号码，即抽样要素；但在目标总体中，家里每个人都应该是一个独立抽样要素。

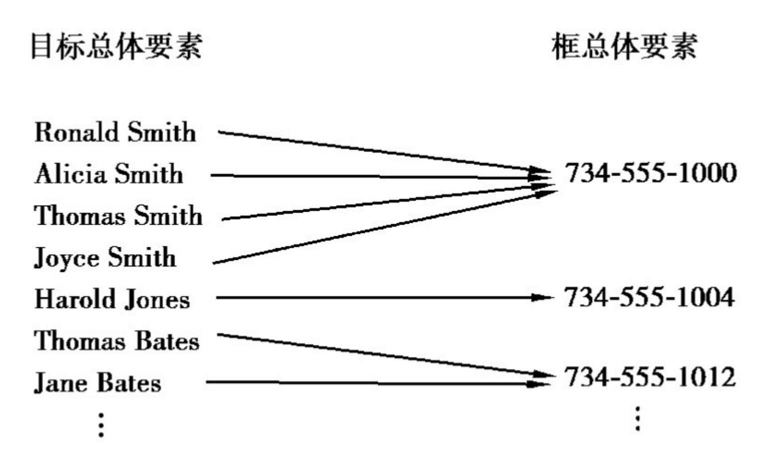


图3.1　一个抽样框要素与目标总体要素的汇聚

应对这种目标总体要素汇聚的方法，就是把样本电话号码所代表的每一个合格个体（汇聚中每个合格个体）都纳入样本。在这种设计中，抽样的概率就是汇聚中所有要素的概率。

汇聚现象提出了一个重要问题，即在汇聚内的二次抽样。第一，在一些情况下，很难从汇聚的每个要素那儿成功地搜集信息。在电话调查中，如果用电话在一个家户访问多于一位对象，就会增加无应答。第二，如果一定要访问一个以上的对象，第一个受访者可能和第二个受访者讨论访问的内容，并因此影响应答。在民意调查中，一个人第一次听到某道访题和在之前已经从完成访问的受访者那里听到过这道访题的应答是有差异的。即使是事实性问题，如果受访者与人讨论过，应答就会发生变化。第三，如果汇聚的规模有差异（如同一个电话号码下的成人个数），控制样本量就成为了难题。要素的样本规

模是汇聚规模的和。除非调查前就知道，否则，就无法直接控制样本规模。

为了避免或减少这类问题，可以从已抽取的框单位（目标总体汇聚）中抽取样本。例如，在电话家户成人调查中，可以从已经抽中的家户中，随机抽取一位成人作为个体样本。所有这些努力，都是为了找到合格个体，也为了消除家户其他人对样本的污染，如此，也可以使得个体层次的样本规模与家户层次的样本规模相等。

在电话或其他家户调查中选择一个受访者，就要建构家户内个人层次的抽样框，从中抽取个人，并对其进行访问。对访员而言，首要工作是通过受访者的应答来搜集数据，但访员很少接受过抽样训练，实际抽样又是数据搜集的一个过程，而且很快，因此在设计上就要保证随机性。具体的讨论，参见4.9节 。

一旦选定样本，在汇聚的抽样中，还有一个问题需要强调：选择的不等概率性。如果所有框要素的被选机会相等，每个框要素只有一次被抽选的机会，与小规模的汇聚相比，则大规模汇聚中合格个体被抽中的概率就要小些。例如，在电话家户调查中，如果一个电话号码代表了2个合格的成人，则其被抽中的概率是1/2；如果是4个合格的个体，则被抽中的机会就只有1/4了。

如此抽样的后果是，合格个体数少的，就会被代表过度，至少对目标总体是如此。换句话说，在目标总体中，从小规模家户中获得的样本比应有的数量大。调查中，如果变量与汇聚的规模之间有关联，那么，对目标总体的估计就不会是无偏的。例如这样的结论：人口规模大的家庭成员成为犯罪受害者的概率大于人口规模小的家庭成员。

为了消除类似的潜在偏差，在分析调查数据时，需要做一些补救。在统计估计时进行加权，使各汇聚中的概率相等是一种方法。第10章介绍了加权和加权后的统计估计。

3.3.4 目标总体要素在抽样框中的重复

在抽样框与目标总体之间多重对应的另一类问题，就是重复。“重复”（duplication）是指一个目标总体要素对应着多个框要素。还是以电话调查为例，其现象表现为一个家户在电话簿上列有多个电话号码。在图3.2中，Tom Clarke是目标总体对象，有两个框要素与之关联，即314-555-9123和314-555-9124两个电话号码。如果一个家户有多个电话号码或申请了多个号码或为此支付了额外的费用，目标总体中的家户就会被多重列入。例如在大学城，没有关系的学生常常会合租一套住房，但却要求一个电话号码。即住户中虽然只有电话号码，但在电话簿上，所有租户的名字都被列在这一个电话号码之下，这就相当于列出了多个电话号码，即让一个目标总体对应了多个框要素。

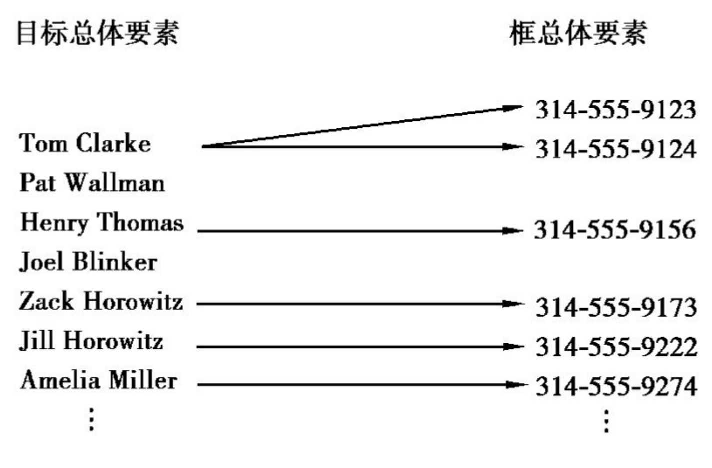


图3.2　目标总体要素在框要素中的重复

这种情况引出的问题与汇聚引出的问题类似。一个目标总体要素对应多个框要素就会造成较高的入选机会，进而形成在样本中的代表过度。如果重复与变量之间有关联，统计估计就会出现偏差。在调查估计中，问题是，是否重复以及重复是否与变量之间有关联，常常是不知道的。

可以采用多种方式处理重复导致的偏差。在抽样之前，我们可以对抽样框进行清理。例如，可以对电子版的电话号码进行清理，消除一个号码多次重复的问题。不过，如果要费劲来清理重复问题，就不经济了。

在抽样甚至调查时，也会发现重复问题。使用一个简单的规则就可以消除问题，即对任何一个入选的样本只选择一个框要素。其他任何的重复，将其都当作外部单位，在抽选时加以忽略。例如，一种方法是，在电话簿中，把重复中的第一个当作合法单位。在与电话家户联系时，也可以询问该家户是否有多个号码或同一号码被重复列在电话簿上。如果是，就可以依据姓进行查找。如果入选的不是该号码，则可以终止访问，将其作为外部单位。

另一种解决方法，和在汇聚中一样，就是加权。如果知道重复的量，就可以根据其在框总体中重复的情形进行加权，使目标总体要素和框总体中要素的被选概率一致。例如，如果在数据搜集阶段发现，一个家户有2条线路，在电话簿上列出了3次，该家户的权重，如果使用电话簿做框，就是1/3；如果使用直拨电话号码做框，就是1/2。

3.3.5　抽样框与目标总体间的复杂对应

多重框单位与多重总体单位之间，也有对应问题。例如电话家户成人调查中，就可能遇到一个家户既有多个成人，也有多个电话号码列在电话簿上。这种多对多的对应问题，就是汇聚与重复的结合体。例如在图3.3中，Schmidt家的3个成人（Leonard，Alice和Virginia）就有2个电话号码（403-444-5912和403-555-5919），即框要素。这就意味着3个目标总体要素要对应到两个抽样框要素上。这种问题的解决方式就是通过对抽样结果进行加权来同时处理汇聚和重复这两个问题。个体层次的事后权重等于用电话号码数除合格的成人数。在这个例子中，Schmidt家的权重就是3/2。如果情况更加复杂，就要运用更加复杂的加权方法来保证两者之间的对应。

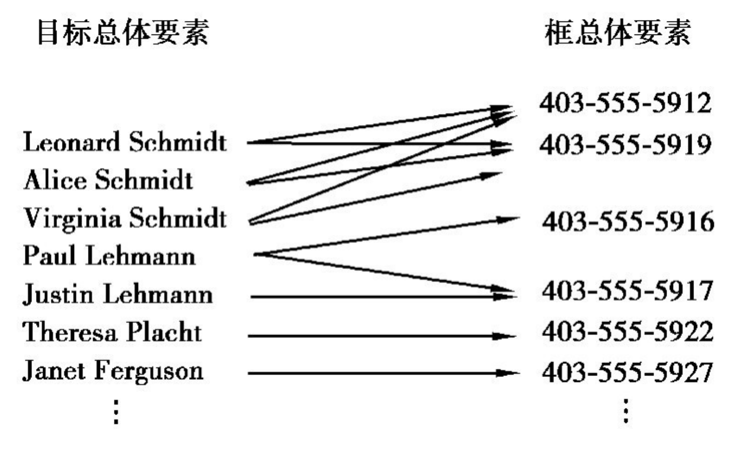


图3.3　目标总体要素与框要素的汇聚与重复

3.4　家户或个人目标总体的替代框

在有上述各种抽样框问题的条件下，这里我们将继续讨论目标总体与抽样框之间问题的表现形式。我们首先看家户和个人目标总体问题，这是社会科学中常用的目标总体。

3.4.1 区域框

在美国，用于家户调查最常见的就是区域框（例如人口普查或县域普查的区域列表）。区域框以地理区域为基础，把个人与区域关联起来，最后落实到居住单位上（依据事实认定准则或法律认定准则）。用这样的框抽取个人，需要经过几个阶段。首先要抽取次级区域，然后列出区域内的地址。如果能列出区域地图或图片，理论上就可以对区域进行全覆盖。如果所列地址不全，就会造成覆盖不足；如果一个人有多个住处，就会造成重复。如果最终的框中，一个地址有多个个人，就会造成汇聚。这些都是全国刑事犯罪受害者调查（NCVS）和全国药物使用与健康调查（NSDUH）遇到的问题。

渐渐的，地址列表在取代或补充先选区域（类似街区）再列表的做法。地址列表的惯常做法是运用美国邮政邮件投递地址列表的办法。这种列表的一个优势是对地址的高覆盖率，劣势是用其他方式获取邮件的地址被忽略了。按照邮政投递的排序方式所面对的一个挑战就是如何把地址与样本区域关联起来。一些商业机构采用匹配方式，如此可提高在区域内抽取居住单位的分层效率。对这种方式的评估还在继续（可参见Innachione，Staab，and Redden，2003）。

3.4.2 用于家户或个人的电话号码框

另一个用于家户人口的框就是与居住单位关联的住户固定电话号码框。这个框大约漏掉了美国住户的20%。还有，少数住户的家里不止一个电话号码，如此就会造成覆盖过度。在用于个体层次的抽样时，

还要清除非家户用电话号码。这就是消费者调查（SOC）和行为风险因素监测系统（BRFSS）遇到的问题。这两项调查都是随机拨号电话调查，用的都是家户固定电话。

在美国，住户电话号码框小于固定电话框。商业机构又从电子版和纸质版的框中，专门摘出了一个框。他们对一般的邮寄调查和调查研究者售卖这样的框。对家户调查而言，用这样的框是很有效率的，因为已经将不再使用的和非住户电话号码清理了。只是，有相当比例的住户电话，不成比例的城市住户电话和暂住人口电话，都不在列表中。这个框也有重复问题，即同一个号码会对应两个不同的人，特别是同一个住户的两个人。

在许多国家，移动电话正在取代固定电话。例如1990年代中期的芬兰，随着移动电话用户数量的快速上升，固定电话用户的数量就开始下降了（Kuusela，Callegaro，and Vehovar，2008）。这样的变化意味着固定电话号码框的失效，因为其中没有包括移动电话。而且，没有包括的主要是年轻人和刚刚从父母家搬出来自己住的人群。

图3.4显示了美国成人中从固定电话转向无线电话或移动电话的速度。在短短4年（2004—2007年）中，成年人中移动电话的使用率从4%增长到了2008年的16.1%。Blumberg和Luke（2008）指出，那些只使用移动电话的人，大多是与没有亲缘关系的人合住的，他们大多是年轻人、收入低、想有独立的住处。

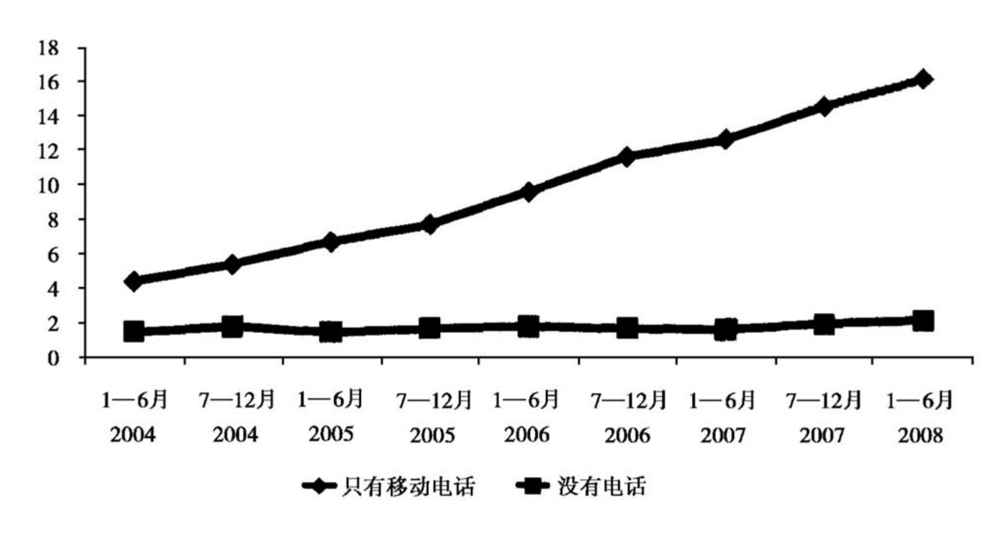


图3.4　美国成人中仅使用移动电话的比例，2004年1月—2008年6月

数据来源：Blumberg and Luke，2008

此外，移动电话与固定电话不同的是，移动电话常常是个人电话而不是家户电话。在这个意义上，使用移动电话建构的抽样框更接近于个人层次的抽样框而不是家户层次的抽样框（容易造成汇聚）。我们注意到，早期对移动电话的研究表明，同一住户中也有合用号码的现象，由此导致的问题更加复杂。

在美国使用移动电话号码框的另一个特征是，移动电话的注册地并不是居住地。这就意味着，如果采用区域抽样，很容易造成对这类人的覆盖不足。

把移动电话框用于电话调查，需要抛开家户框以及不把家户作为抽样单位。在现实条件下，无论是固定电话框还是移动电话框，都有汇聚和重复的问题。在美国，运用移动电话的最大难题就是号码不连续。在这样的条件下，就要使用不同的号段进行抽样，同时也使用其他号码如固定电话号码抽样。如果是这样，会遇到有人既有移动电话也有固定电话的情形，即同时出现在两个框中。有些设计，试图通过

加权来处理这类重复问题；有些则试图通过以个体为基础，排除重复的号码。两种方法各有利弊。

运用手机号码框还有一些其他的问题，可以参见www.aapor.org/uploads/Final_APPOR_Cell_phone_TF_report_041208.pdf。其中包括了第6章讨论的无应答和第5章讨论的测量。

3.4.3 对一般总体进行互联网调查的框

随着互联网调查的发展，人们开始认真琢磨是否可以建立一个家户层次的电子邮件框。但电子邮件框常会漏掉家户的大部分人口。也有重复问题，因为一个人可以有多个电子邮件地址；也有汇聚问题，因为许多人可以共用一个电子邮件地址（例如smithfamily@aol.com）。对互联网调查而言，非标准化的电子邮件地址也会和直拨电话调查一样产生混乱。当然，电子邮件不是互联网调查的唯一方式，它还可以用于邀请一些人直接访问某个调查用的网站。

由于缺乏通用的电子邮件地址框，主动加入的互联网追踪（opt-in internet panel）（或可及性追踪）常试图累积大量的电子邮件地址，并稍后给这些地址发送调查邀请。人们常常说自己已经累积到推及全部成人总体的样本地址，但需要清楚的是，即使如此，也是个体层次。

第一，这类互联网追踪调查企业的标准不是全覆盖，而是累积到与调查相关的、多样性足够的地址。对于家户层级的产品而言，多样性包括了收入、家户规模、族群亚文化、年龄、生活方式等。对公众

观点而言，多样性则包括了收入、年龄、性别、族群、政治认同、社会参与等。除了互联网调查企业以外，还可以从其他组织如互联网零售组织、会员组织等获得电子邮件地址，并对其进行识别，了解其接受调查的意愿。一旦达到了委托调查机构要求的多样性和规模，就可以将其用于调查了。只是，在真正使用时，要对每个地址对应的群体属性做甄别。

第二，在互联网追踪调查中，所谓框的概念就不存在了，取而代之的是直接抽样和调查。互联网上人是通过口号（如“回答问题就赚钱”）或弹出窗口汇聚到调查中来的。因此，这类调查的任务就是从足够多样性和大的群体里获得数据。

第三，互联网调查常常是在那些愿意接受调查的群体中做调查，只要对方愿意接受调查，调查机构就会向其发送各种问卷。如果要观察总体的动态，就需要注意这样的做法所产生的系统偏差。市场研究机构在制订一些规则，用于剔除那些不可用的自愿加入的地址，以及不回答问题或虚假的受访者（参见www.esomar.org/index-php/26-question.html）。

简而言之，由于没有与个体关联的统一的电子邮件地址框，一些调查就干脆忽略了对框的建构。但如果没有一个定义良好的抽样框，就完全无法估计覆盖性偏差了。

3.5　其他一般目标总体的样本框问题

这一节将简要介绍消费者调查、就业者调查、组织调查、事件调查以及稀有总体调查的抽样框问题。

3.5.1　客户、雇员或组织成员

大多数对客户、雇员或组织成员的调查采用的都是列表框。有时候，列表框是个体记录的电子文档，另一些时候，也可能是物理性记录。这些记录系统常常有可以预料的覆盖性问题。覆盖不足问题来自于文档更新的不及时。如果在记录更新之前要经过很多个行政步骤，类似于就业者和客户的信息，就可能造成更新不及时。

同样，列表中也可能包含不合格要素（ineligible elements），特别是当个体离开了组织，但列表信息没有及时更新时。例如，在一个客户列表中，一些客户可能很早就认为自己已经和组织完成了最后的交易，且不再是组织客户了，但列表中依然还列着他们。此外，是否把一些人算作组织的成员，也有模糊不清的地方。在商业组织雇员调查中，如何看待组织与雇员之间的合同？尽管他们日复一日地在工作，但他们也许是作为另一个组织的成员在为其提供服务。

客户或成员框内也可能出现重复问题，但相对于家户而言，会少一些。在客户框中，如果记录是基于组织与客户之间交易的，则不可避免地会出现重复问题。只要是与组织发生过交易的成员，都会有记录，这样，调查研究者就需要细致地思考，看看目标总体是人（即客户）还是交易，或者两者都是。

渐渐地，制作工作或成员组织抽样框开始利用组织内部互联网沟通的优势了。如果雇员或成员有且只有一个组织的电邮地址，则可以制作几近完美的抽样框。但如果电子邮件地址是一个职位的而不是个人的，问题就会来了，因为一个人可能既有职位的地址，也有个人的地址。不管怎样，这类框避免了上一节讨论的互联网调查中提到的问题。

调查研究者们尝试着如何以及为何可以将一种框作为另一种框的替代。例如，在医院，依据薪酬表做的列表可能会漏掉义工和临时工，但如果依据安全系统（如门禁系统——译者注）的列表，也许可以避免这个问题。知道如何以及为何要更新和修正列表，非常重要。例如，按月支付薪酬的列表较之于按周或按天支付的列表，更新就慢。对框而言，涉及临时缺席的更新过程，就牵涉到了复杂的覆盖性问题。如果一个雇员是编外的医疗人员，他或她是否应该列入抽样框呢？目标总体是否要包括这类人员？所有这些问题，都要根据调查的具体情况来决定。

3.5.2　组织

组织总体是多样化的。例如教堂、商业组织、农场、医院、门诊、学校、慈善、政府部门以及民间组织。这类组织的抽样框，常常是组织单位的列表。就所有的组织总体类型而言，商业组织常常是调查研究的对象。

商业组织总体有着特殊的建框问题。第一，特别突出的特征就是，商业组织的总体常常在规模上差别很大。如果选择的是软件商总

体，微软（年销售额达200亿美元）和街头商店（年销售额5 000美元）都在其中。许多商业调查的变量都与商业组织的规模有关（例如，当前就业统计调查就是估计产业中的总就业人口）。因此，在商业组织调查中，覆盖性问题更多强调的是不要漏掉最大型的组织。

第二，商业组织的动态性很强。小型组织的开业与关张常常非常迅速。较大组织的并购或合并也是常有的事情（例如惠普并购康柏，并成为了一个企业）。商业组织拆分也是常有的事儿（例如福特汽车公司把零部件拆分出来成立一个独立的维斯顿公司）。这就意味着要经常更新抽样框，以便在先前抽样框的基础上保持较好的覆盖性。

第三，商业组织还要区分法律上的组织和地域上的组织。多部门和多地点的组织是很常见的（例如全球有超过3万家麦当劳营业点，但只有一个总部）。因此，对商业组织的调查可以调查企业、法律意义上的实体或营业场所。有些合法的商业组织甚至没有营业地点（如咨询组织的员工在自己家里办公）。有时候，一个场所也会有多个组织，但老板可能只有一个人。

除了商业组织总体以外，其他组织总体也有类似的特征，或多或少比较相似。这些特征要求研究者认真探讨组织规模的变异性，总体的动态，以及法律和区域上差异。

3.5.3　事件

有时候，调查对象的总体是事件。调查中，有许多类型的事件：产品或服务的购买、婚姻、怀孕、出生、失业时间、抑郁间隔、穿越

隔离带的汽车，或刑事犯罪的受害情况（如全国刑事犯罪受害者调查，NCVS）。

通常，对事件的调查都从个体框开始。一个个体要么没有，要么经历过某个事件。有些人会经历多个事件（如多次购买）并因此落入事件要素的汇聚中。NCVS就是把受害作为事件研究的。首先用受害者建构一个框，每个人都是一个潜在的受害事件汇集。NCVS测量之前6个月的每个事件，然后统计受害事件的特征。

用于事件抽样的另一个逻辑性框总体就是时间单位总体。例如，对一年内到动物园的参观抽样。调查的目的也许是了解到动物园的目的，参观的时长，动物园最有意思的部分，或者最没有意思的部分。制作抽样框的一种方式就是把每个参观与时点结合起来，如出园的时间。按照这样方法，所有的参观都会对应到相应的时点上，而且每次参访只有一个时点。如果某个研究要抽样，就可以选择一个次级时点（如5分钟一个间隔），如果抽到了某个事件间隔作为样本，就在其离开动物园5分钟内对其进行访问。

有时候，运用时间的调查（即了解总体群体在一定的时间范围内做什么的调查）会用电子随机时间鸣叫器进行时间选择。当鸣叫器鸣叫时，就调查这个时点的受访者在做什么（如在办公室工作、看电视、购物）（参见Csikszentmihalyi and Csikszentmihalyi，1988；Larson and Richard，1994）。

事件调查可能同时包括多个总体。在统计上，对事件总体要做研究，对涉及事件的人的总体也要做研究。如果同时包含这两个目的，就会出现多种汇聚和重复问题。在对购车者的研究中，研究到底是一个家庭的购买事件、购车的人（法律上的车主）、所有家庭成员，还

是该车指定的驾驶者？NCVS产出的是经历了刑事犯罪受害的家户的比例（以家户为基础）和户内成员的比例（以家户内人口为基础）。仔细探究要估计的变量对选择事件研究的目标总体和框总体非常重要。

3.5.4　稀有总体

“稀有总体 ”（rare population）是指研究者有兴趣的小型目标群体。有时候，说稀有，并不是指绝对规模，而是指框所覆盖的相对规模。例如，在美国，福利的受益总体，在30 500万人中有750万人属于福利受益群体，由于其只占总人口的3%，因此，可被看作稀有总体。如果把稀有总体作为目标总体，对制作抽样框而言，就有诸多问题。

对稀有总体而言，建构抽样框有两个基本方法。第一种方法就是列出可以列出的稀有总体要素。例如，我们可以从福利办公室直接得到所有福利受益人的名单（尽管这些信息有可能是保密信息）。有时候，可能不会有覆盖完整的一份名单，但可以用多份名单合并（参见3.7.3 关于多框设计的讨论）。第二种方法更加普通，建构一个可以包含稀有总体且可以把稀有总体筛选出来的框。例如，可以从家户总体中筛选出福利受益家庭。如果稀有总体的所有要素是其上一级稍大的框总体成员，那么就可能获得对稀有总体的全覆盖（取决于从样本单元中筛选出稀有总体的成本）。

3.6　覆盖性误差

针对3.3讨论的许多抽样框问题，有不少解决的方法，但所有的解决方法都不能总是消除覆盖性误差。在调查中，覆盖不足是一个难题，也是覆盖性误差最重要的来源。非常重要的是，覆盖性误差是抽样调查以及通过调查进行估计的一个特点。对一项调查而言，一种统计可能有较大的覆盖误差，另一种统计则不一定受到同样的覆盖性误差困扰。在调查方法中，覆盖不足、重复、汇聚以及其他问题，都是抽样框问题。覆盖性误差，则是这些问题在统计上的效应。

在简单统计如样本平均值中，覆盖性误差的本质已经在2.3.4节中进行了讨论。对预测平均值而言，覆盖性偏差可以表示为：

$$\overline{Y}_C - \overline{Y} = \frac{U}{N}(\overline{Y}_C - \overline{Y}_U)$$

式中，$\overline{Y}$ 是总体的均值，$\overline{Y}_C$ 是被样本框覆盖的总体的均值，$\overline{Y}_U$ 是未被样本框覆盖的总体的均值，相应的，U 是所有合格的但却未被样本框纳入的总数（未被覆盖的数），N 是样本总体的总数，C 是样本框内合格的总数（被覆盖的数）。这样，没有覆盖的N-C 的误差，就是未被覆盖的比例与覆盖的和未覆盖的均值差的函数。

调查（与样本量无关）只能估计覆盖了$\overline{Y}_C$ 的均值U 中没有被覆盖的程度，或覆盖与没有覆盖之间的实质性差异，决定了偏差的大小或覆盖误差。每一层级没有覆盖的比例会有差异。因此，总样本误差会比某个子总体的误差要大。此外，由于覆盖性误差取决于覆盖的与没有覆盖的估计值之间的偏差，因此，一种统计与另一种统计之间的覆盖性误差会不一样，即使运用一样的子样本。

3.7 减少覆盖不足

针对一般的覆盖性问题如抽样框中的重复、汇聚、多对多的映射、覆盖不足，以及外部要素等问题的解决之道，已经在3.3节中讨论过了。然而，专门针对覆盖不足以及后续的覆盖不足误差，还没有进行过详细讨论。事实上，有一些程序就是用来改善覆盖性的，包括专门用于减少误差的抽样框补充设计。

3.7.1 半开放距

如果抽样框稍稍有些过时，或者除针对某类单元以外，覆盖性不错，就可以通过在调查前或调查中更新列表的方式进行弥补。如果有一个列表逻辑，就有可能通过比较两个列表之间缺失的部分来修复抽样框。

例如家户调查中的地址或住宅列表（参见图3.5）。这类列表很容易产生过时或缺损。但如果审核仔细，也可以将缺损的补回到列表中。由于地址列表依据的是特定的地理位置顺序，因此对特定的框要素而言，是可以采用追加的方式将缺损的单元补回去的，而不需要更新整个列表。也就是说，在抽样以后或调查中，是可以更新列表的。

序号	地　址	选　择
1	101 Elm Street	
2	103 Elm Street，Apt. 1	
3	103 Elm Street，Apt. 2	
4	107 Elm Street	是
5	111 Elm Street	
6	302 Oak Street	
7	308 Oak Street	
⋮		

图3.5　家户调查的街区区域地址列表

更新的一种方式被称之为“半开放距　”（half-open interval）。回想图3.5，即一个街区的地址列表。图3.6展示了街区的概要，即可以看到住房的地理分布。假设我们抽中了这个街区，即Elm街107号。从区域框的视角，107号不是一个物理结构，而是一个地理位置，这个位置是独享的，也不包括下一个位置如111号。在数学上，依据数学上的集合论，采用半开放距就可以让每个地址都出现在列表上。从107号（距的封止点）起始往后延伸，但不包括111号（距的开放点）。

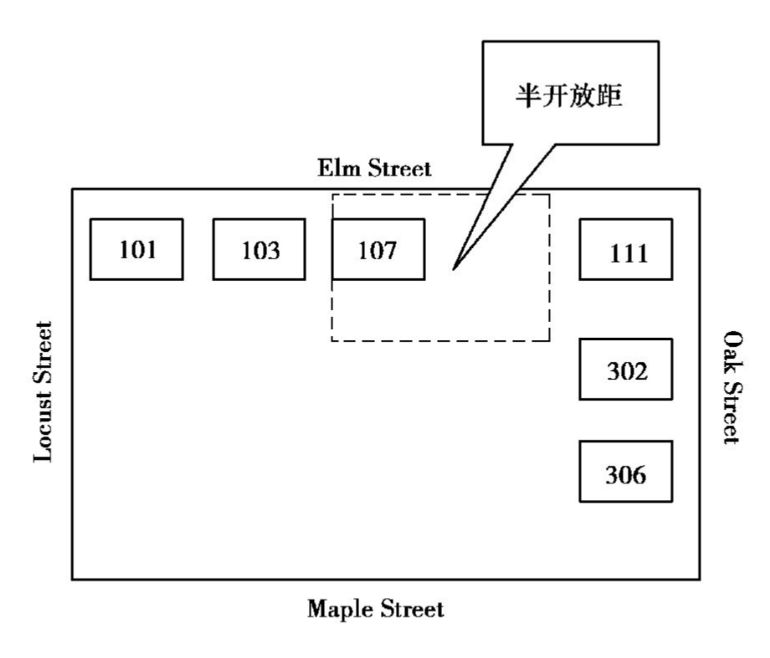


图3.6 样本街区住址概要

当访员到达样本地址时，他不仅发现半开放距关联的是107号，也还要看在这个距中是否有新建的单元或任何缺损的单元。如果发现有，访员就可以将其加到列表中，将其作为样本地址，并对所有地址进行访问。在半开放距中的所有地址，与之间的样本地址都有相同的入选概率。这样，在调查中，缺损的或新进的单元，就会自动地列入样本框中。

有时候，在半开放距中发现的新址数量过大，以至于访员无力继续。例如，如果在107号和111号之间新建了一座公寓楼，有12套公寓，访员就要访问13户，而不是1户。在这种情况下，可能就要对新出现的地址进行二次抽样，以减少其因汇聚效应对样本量和访问工作量的影响。就像在汇聚中的二次抽样一样，由于入选概率的差异，所以一定要进行加权（参见第10章 的权重和加权估计）。在连续性调查中，也可以将这类加进来的部分作为一个单独的层进行处理（参见Kish and Heps，1959），用不同的概率抽取缺损的或新加的样本。

同样的思路，也可以用于对样本框覆盖性的核查。例如，一项调查儿童就读公立学校的地址列表，调查时，就可以更新样本家户样本儿童的年龄，以及缺损或新生儿童以及他们的年龄。

3.7.2　倍增抽样

半开放距概念通过在抽样或调查过程中搜集信息来补充既有的框。用于对框进行补充的还有网络抽样，通常也称之为“倍增抽样 ”（multiplicity sampling）。选择一个样本，加上定义清晰的样本涉及的网络。

例如，家户调查中抽中了一位成人，就可以询问其所有或者部分的兄弟姐妹，这就形成了一个可以搜集信息的样本网络。当然，由于存在重叠，网络成员有多重被选机会。网络的规模决定了重叠的量。如果样本成员说他有两位活着的兄弟姐妹，那么网络的规模就是3，估计时的权重就是1/3，由此以减少因网络规模所造成的概率差异。倍增抽样和加权方法（Sirken，1970）在稀有总体中用于增加样本规模进而获得稀有样本，如某种疾病。当然，使用这种方法时，需要考虑网络中个人的隐私问题。此外，应答误差（参见第7章 ），如落下了某个兄弟姐妹，或纳入了某个非兄弟姐妹，或对网络成员的特征误报，对调查针对的问题而言，就会造成网络定义和覆盖性的误差。

“滚雪球抽样”则是另一个密切相关的、补充抽样框的、非概率方法。假设在调查中找到了一个稀有样本，如失明，通常他们对同类都有所了解。我们就可以询问样本，了解其他失明的人，他们告知的其他人就可以追加到样本中。滚雪球抽样就是这样通过样本的网络来

累积样本的。告知的误差，与网络没有任何联系的个体，以及界定不周延的网络，造成了滚雪球抽样实际应用的困难，渐渐地也不再是概率样本。

尽管在理论上倍增抽样为解决抽样框问题提供了解决之道，但在实践中，仍然有许多需要解决的问题。包括在网络告知中的测量误差，因网络告知不完全所造成的无应答误差，以及倍增估计中的变异性波动。

3.7.3　多框设计

使用多框抽样 （multiple frame sampling），在许多环境下，会降低覆盖性误差。在主框几乎覆盖了目标总体的条件下，辅助框则可以更好地专门覆盖主框没有覆盖或覆盖不好的要素。例如，一个过时的住户列表，就可以辅之以从政府负责样本区域建设部门拿到的新建住宅列表。另一个例子就是，对流动房屋的覆盖常常会比较差。专门建立一个流动房屋的补充列表，就可以更好地解决主框对其覆盖不足的问题。

许多情况下，辅助框常会完整地覆盖总体的某个特定部分。但在更多的情形下，辅助框会与主框发生重叠。在这种情况下，就要采用多框抽样及其估计程序，以解决选择的不等概率问题，甚至可能因此而提高估计精度。

例如，假设使用一个完整的美国家户电话号码框，采用随机拨号的方式（RDD）进行电话调查。理论上，RDD包括了全美的家户电话，但却不包括2%左右没有电话的家户。图3.7展示了电话框作为家户单元

区域框的影子框的情形。要解决覆盖不足问题，就可以采用辅助框，即区域框。只是，区域框样本需要到户，比使用电话花费更多。总之，两个框就可以对家户完全覆盖。就全国刑事犯罪受害者调查（NCVS）而言，Lepkowski和Groves（1986）探讨了RDD和区域框的成本与误差问题。他们发现，对于硬预算约束而言，在大多数样本来自于电话框（基于模拟的，包括覆盖性估计，无应答，以及测量误差）时，在统计上会获得较低的均方差。

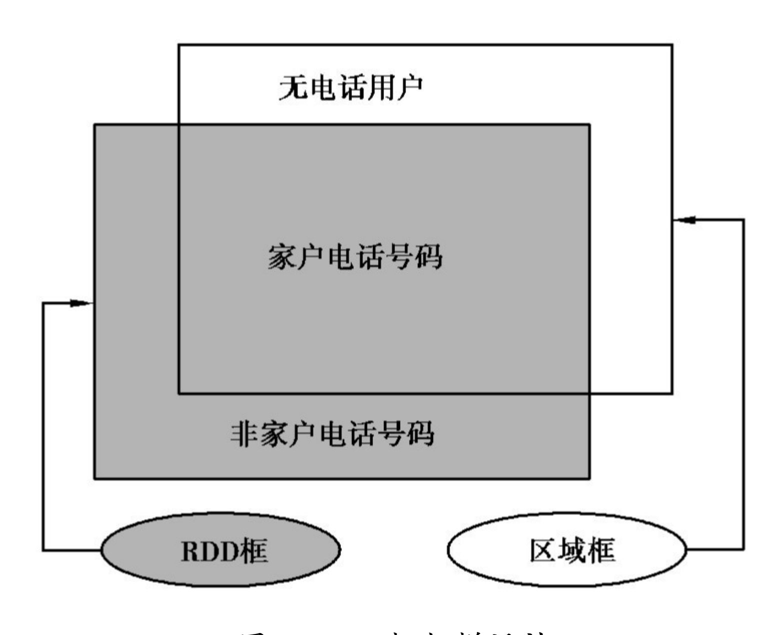


图3.7　双框抽样设计

但两个框是有重叠的，每一类都包括了有电话的家户。这种包括了两种框的数据集，实际上混合了来自两个框的数据。清楚的是，在这种设计下，有电话的家户被过度代表了，因为两个框里都有它们。

对重叠和过度代表的问题，有许多解决方法。第一种方法是对区域性样本进行过滤。在进门访问开始之前，由访员了解该户是否有固定电话，即是否已经在电话框里了。如果已经在电话框里了，那么就不纳入样本，也不做访问。运用这个过程，就会消除重叠问题，这样，双重框也能完整地覆盖家户。

第二种解决方法是访问两个框中的所有样本家户，并确定两个框的入选机会。没有重叠的、没有机会从电话框中入选的，就只有一个机会，即从区域框中入选。有电话的家户，就有两个机会入选，一是从电话框入选，二是从区域框入选。因此，它们的入选机会是$P_{RDD}+P_{area}-P_{RDD}\times P_{area}$，这里$P_{RDD}$和$P_{area}$分别代表电话号码框和区域框的入选机会。不管使用的是哪个框，对非电话框样本而言，其权重为$1/P_{area}$；对电话框样本而言，其权重为1/（$P_{RDD}+P_{area}-P_{RDD}\times P_{area}$）。

第三种解决方法是Hardley（1962）等提出来的。他们建议将重叠的部分用作更加有效的估计工具，即将双重框用作一个没有重叠的区域集，然后用两个框一起来获得目标总体的估计值。如图3.7所示，有三个区域：无电话家户（*Non-tel*），有电话家户（*RDD-tel*），区域框有电话家户（*area-tel*）。*RDD-tel*和*area-tel*一起，运用混合参数可以构成了一个数学上最大化的精确估计（譬如说，均值）。有电话和无电话一起，并运用权重，即电话家户在目标总体中的比例，譬如W_{tel}。如此，双重框对这个特定样本的估计就是：

$$\bar{y}=(1-W_{tel})P_{non\text{-}tel}+W_{tel}[\theta P_{RDD\text{-}tel}+(1-\theta)P_{area\text{-}tel}]$$

式中，θ是用于最大化估计精度的混合参数。

图3.7是多框抽样的一个特例。但其方法可以用于更加复杂的情形，如3框、4框；框越多，产生的重叠区域就越多。即使在双框抽样中，也至少有4个区域：框1；框2，框1与框2交叉；框2与框1交叉。尽管后两个是两个独立框的交集，如果真的用于抽样，就会影响到参数

估计，因此，要把它们分开看。在农业调查中，常会看到这类设计。假定要针对特定的牲畜如奶牛抽样，假定在国家的农业部门有奶牛农户的清单，但这个清单有些过时了，其中一些农户没有再养奶牛了，同时，清单中也不包括养着奶牛的小农户。第二个区域框是用来针对所有农户抽样的。这就有了4个域：单纯的农户清单，单纯的区域清单，包含在区域框中的农户框，包含在农户框中的区域框。同样，要解决重叠问题，就需要过滤、加权，或进行多框估计。

最后一个例子涉及最近大家有兴趣的网上调查设计。假定商业公司能够提供电子邮件地址清单。将其用于自访问卷的调查将会非常节省经费。但是，会出现外部要素 （foreign element）（不再使用的地址，不再是合格对象）和不能对合格群体完整覆盖的问题。这就需要使用电话号码框作为补充框。虽然费用会上升，但却提高了对合格群体的覆盖性。可以两边抽样，两边调查，两边估计，然后把结果进行合并。对这种混合执行方式，现在有很多的研究（参见5.4节 ）。

3.7.4　通过纳入更多适用要素增加覆盖性

如果调查的统计产出是个体层次的，则提高覆盖性的最后步骤主要针对样本家户中的个体。一旦识别了样本家户，访员就要把居住在家户的所有成员列出来。常见的情形是，如果某些人不是常在家居住，或者提供信息的人理解的家人与调查定义的家人不同时，就很容易在列表时漏掉。

Tourangeau，Shapiro和Ernst（1997）以及Martin（1999）都研究了在询问到哪些人居住在家里的时候所出现的情形。同样的提问是“都有谁住在这儿？”另外的问题是，昨天晚上，都有谁吃住在这儿？或居住的房间？或房子的钥匙？谁会在这儿收信？这些人通常会住在这里，但也会短暂离开（参见文本框 ）。这些问题，会减少在列表时出现遗漏的情形。

Tourangeau，Shapiro，Kearney，Ernst（1997）和Martin（1999）论家户问题

有两项研究讨论为什么在家户成员列表时会漏掉成员。

研究设计：Tourangeau等人针对三种程序设计了一个实验：询问居住其中每个人的名字；询问前一晚居住其中的每个人的名字；询问前一晚居住其中的每个人的俗名或昵称。问完之后，确认列出的每个人是否符合定义。在3个城区的49个街区，访问了644个居住单元。Martin的研究则调查了区域中的999个概率性居住单元。询问在过去的2个月任何与该单元有关系的个人名单，在此基础上进行二次抽样，确认其是否符合定义。在两项研究中，在做出列表后，都对列出人员的名单是否符合“居住”的定义进行了进一步的确认。

研究发现：Turangeau等人的研究发现，只有询问俗名或昵称的询问获得更多的成员，他们认为隐瞒某些人员是导致覆盖不足的主要影响因素。Martin的研究发现，常有前后不一致地报告无关人员的情形，如一周或之前就已经离开了，或者对家里没有经济责

献。Martin的结论是，家里提供信息的人对家人的定义与调查定义不一致，导致了覆盖不足。

研究局限：两项研究都无法给出家户成员的真值。两者都认为后面的追问会更加接近真实。

研究影响：提供了家户清单错误的规模。两者都展示了家户定义的复杂性以及不乐意报告非常在人员所导致的覆盖不足问题。

接下来的步骤，是依据调查的定义，确认列出来的人是否为合格的家户成员。经过这样的提问，就能够把在其他地方还有住房的人甄别出来。

这种宽进严出的策略，能够把应该包括的样本纳入进来。不过，这样做的缺点就是需要更多的时间和提问来识别家户中的合格对象。在许多情形下，这些问题是访员首先要问的问题。因此，争取家户信息提供者的支持尤为重要，也很少有替代方案可用。

3.8　小结

在调查设计中，目标总体、样本框，以及覆盖性都是重要的议题，因为它直接影响到依据调查数据进行的统计推论。当把框总体和目标总体进行比较时就会发现这类问题，这也是解决这类问题的惯常方法。当然，如此，也不一定就完全解决了覆盖性问题。

调查中，覆盖性误差独立于抽样步骤。抽样是在框总体中抽取样本。样本不可能优于其框。下一章将讨论的就是抽样。前提假设是，

对总体参数的估计而言，这里讨论的覆盖性误差和框总体问题与如何抽取样本是完全不同的问题。

关键词

区域框（area frame）

区域概率样本（area probability sample）

汇聚（clustering）

覆盖性（coverage）

重复（duplication）

事实认定准则（de facto residence rule）

法律认定准则（de jure residence rule）

要素（element）

外部要素（foreign element）

外部单位（foreign unit）

半开放距（half-open interval）

家户（household）

居住单元（housing unit）

不合格要素（ineligible element）

不合格单位（inelibible unit）

多重对应（multiple mappings）

倍增抽样（multiplicity sampling）

多框抽样（multiple frame sampling）

主动加入的互联网追踪（opt-in internet panel）

拨号调查（random digit dialing，RDD）

稀有总体（rare population）

抽样框（sampling frame）

调查总体（survey population）

目标总体（target population）

覆盖不足（undercoverage）

常住成员（usual residence）

进一步阅读资料

Kish，L.（1965），*Suwey Sampling*，Chapter 11，Section 13.3，New York：Wiley.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys* , Chapters 3-5, New York: Wiley.

Levy, P. and Lemeshow, S. (2008), *Sampling of Populations* : *Methods and Applications* , 4th Edition, New York: Wiley.

作业

---

1. 运用教材中6个调查（参见第1章 ）中的一个，描述目标总体或抽样框的替代定义，用以在调查结果推论中消除覆盖性误差。（需要说明你头脑中调查估计参数、目标总体、样本框）

2. 如果将一个针对美国家户成人的区域概率样本面访调查转化为一个互联网调查，并做同样的统计估计，指出你关注的3个问题。

3. 如果要通过电话调查描述所有没有家户的成人，指出不会出现覆盖性问题的两个条件（不管是不是现实）。

4. 假设你对360平方英里、3个县域内的农业工人有兴趣，但你却没有农业工人的列表，为此采用网格套地图的方法，并准备从网格中抽取一平方英里的样本。识别使用这样的方法抽样可能遇到的3个问题。

5. 距上次人口普查已经5年，你要用电话号码框进行家户调查。假设抽取的是家户电话号码，且要求访员访问家中对家户成员健康状况最了解的成员。调查结束后，有人提出通过比较上次普查的人口分布（如年龄、性别、种族、民族等）和调查中提供信息者的分布，对你的调查进行评估。你将如何评价这个提议？

6. 假设你为总统候选人工作。她给了你一份互联网上自愿填答的结果，并证明她领先于主要对手。互联网上的自愿填答者来自支付了信用卡高额年费和有巨大收益（如产品和服务的折扣）的群体。这份调查结果相对于使用电话抽样和电话调查而言，更加有利于你的候选人。她要求你对两份调查结果进行评估，以及如何向受访者提问什么问题。你会如何对她说？（提出2～4个评估点）

7. 在世界上的大多数国家，手机号码与固定电话有明显的区别。如果正如本章所表述的区别，创建一个参数集，用于电话调查情境下无法覆盖样本所产生的覆盖性误差。

8. 假设你要对过去两年去过门诊的患者进行抽样。可以提供给你做抽样框的资料是档案室的纸质记录，包括患者及其家人的造访记录，还有记录有每个人信息的表格。根据本章讨论的每一种框的缺点，列举这种情况下建构抽样框的缺点。

---

(1) 也有译作“单元”的。elements的基本意思是最小单位的构成物，“单元”与“单位”之间，因“位”和“元”原本同义，很难区分；故这里译作“要素”，意指构成物。

4　抽样设计和抽样误差

4.1　导言

从一个抽样框中选择要调查的样本是调查过程的重要一环。一项调查要有完美设计的问卷，培训严格、积极性高的访员，出色的实地督导和管理，与数据类型匹配的搜集方式，以及良好构思的数据整理方案，但是，如果样本选择随意或主观，也很难达成调查的目标，即对总体进行推断。

1998年的秋天，美国地理学会（NGS）启动了基于互联网的“调查2000”项目，希望了解人们看电影、参观博物馆以及读书的频率。调查者在NGS网站和NGS的期刊上刊登广告，敦促人们参与调查。最终有80 000人访问了网站，50 000人完成了问卷（Witte，Amoroso，and Howard，2000）。

稍早时候，1997年，美国艺术品捐赠协会主持了“艺术中的公众参与调查”（SPPA）（National Endowment for the Arts，1998）。SPPA是一项基于随机产生的电话号码的家户电话调查。调查的应答率约为55%，获得了12 000个成人的应答。问卷中测量了不少与NGS调查一样的行为。

SPPA的调查结果基于严谨一些的抽样方法，与NGS基于被访者自我选择调查的结果大相径庭。例如，NGS的调查中有60%的人说自己在过去的12个月中看过演出，而在SPPA的调查中，只有25%的人看过音乐演

出，16%的人看过非音乐演出。同样，在NGS的调查中，有77%的人说自己到访过艺术博物馆或画廊，而SPPA的调查结果只有35%。

调查中，获得样本的方法会直接影响调查结果。NGS调查中自我筛选的样本，也许原本就对文化活动更有兴趣（Couper，2000）。

与NGS调查形成对照的是消费者调查（SOC）。SOC的样本，除了阿拉斯加和夏威夷以外，随机地选自于所有有电话的家户。“随机” 选择（random selection）或“机会”选择，意味着所有人为的或非人为的因素，在选择过程中都被剔除了。在样本选择中使用的随机数，其出现，无论是一位数还是多位数，都是非顺序性的。同时，选择也要控制样本在地域上的分布与目标总体一致。随机抽样机制和地域控制可以避免样本出现偏差如收入偏高、少数族裔偏少、女性偏多以及任何与总体分布不一致的情形。

简单地说，在精心设计的抽样中，有三个主要特征：

1）一个总体要素的列表或列表的组合（第3章讨论的抽样框）。

2）从列表中随机选择样本。

3）运用一些机制，保证样本可以代表总体。

特别重要的是，仅有随机选择并不能保证样本的代表性。例如，样本的确是从48个州的家户电话号码中随机抽选的，但抽到的都是城镇地区的家户。虽说是随机选取的，却不能代表总体。因此，一个好的抽样设计，既要保证随机性，也要保证代表性。

如果使用随机数表在抽样框中进行抽样，那么所得到的样本就被称为“概率样本 ”（probability sample）。所谓概率样本，是指抽样框中的每一个要素都有一个非零的被抽中的概率。被抽中的概率也不一定一样。例如，设计者为了对某个小群体进行单独估计，或许要在总体中有一个过度被代表样本集，譬如70岁及以上的老人。过度被代表是指某个群体相比其他的群体而言有着更高的入选样本的机会。也就是说，虽然采用了概率抽样，也不是每个要素入选样本的概论都是一样的。

2.3节已经区分了固定误差或偏差以及变量误差或方差。抽样会使得两类误差增大。抽样偏差和抽样方差指的是并非抽样框的所有要素都被测量过了。如果因为抽样设计的原因使得一些要素系统性地被忽略了，那么，在调查统计中，就会出现“抽样偏差 ”（sampling bias）。例如，如果一些人在抽样框中的备选机会是零，但他们却有着特异的被研究特征，如此，由被选中的样本产生的估计值就会要么偏高，要么偏低。即使是所有抽样框的要素都有被选中的机会，用同一个抽样设计，也会获得不同的样本，由此产生的估计值也会不同。如此变异性，正是抽样统计中“抽样方差 ”（sampling variance）的基础。我们会不断地用“精度 ”（precision）来表示方差的程度。

4.2　样本值和估计值

并不是所有用于调查的样本都是采用概率方法获得的。许多调查会使用偶遇样本或目标样本。例如，商场入口处针对每个进入商场顾客的调查，调查人员可以一直守在入口处进行访问，直到完成预设的

样本数为止。偶遇或方便抽样方法都有一个共同的弱点：无法在理论上将其推论到较大的框总体。不过，概率样本可以用于统计推论，即用样本数据在一定的置信区间，推论到总体。

要理解抽样，就需要对调查时观察不到的一些重要概念有理解。图4.1显示了变量Y的值在框总体的分布。以消费者调查（SOC）为例，其框总体为所有有电话家户里的成年人。不过，我们从未见过其分布，也从未对其有充分了解。调查的目的，就是了解。其分布有一个均值，我们用大写的字母$\overline{Y}$表示。正因为我们对其不了解，所以要通过调查来对其进行估计。每一个Y值即Y_i构成的分布，我们用S^2表示，即总体要素的方差。

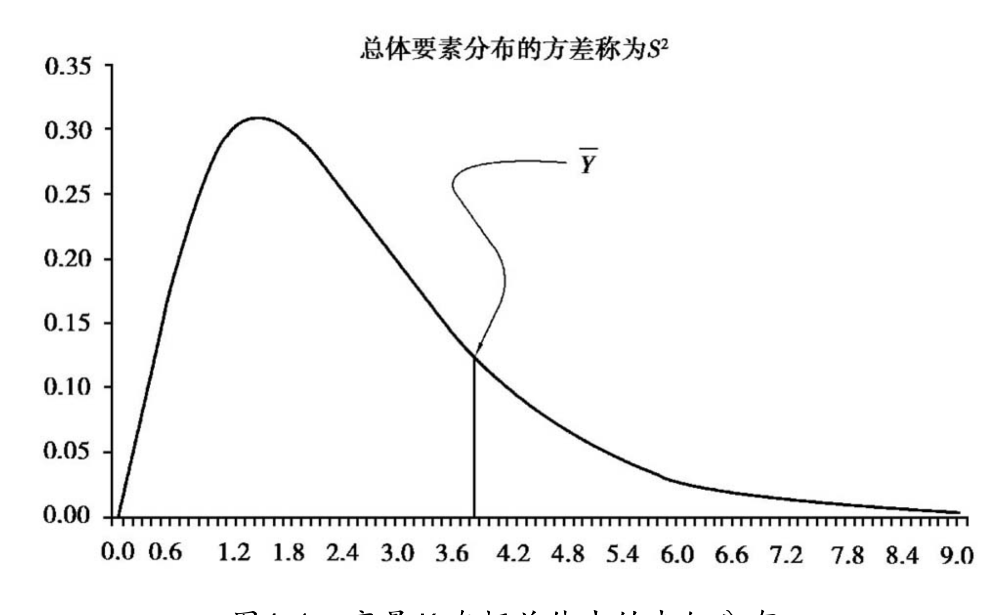


图4.1　变量Y在框总体中的未知分布

请记住，在本章，大写字母总是用来标识框总体的特征，也是我们不知道的特征。

图4.2是第2章用过的一幅图，表示用样本估计总体的过程。当采用概率抽样方法时，要抽取的不是一个样本，而是许多样本。

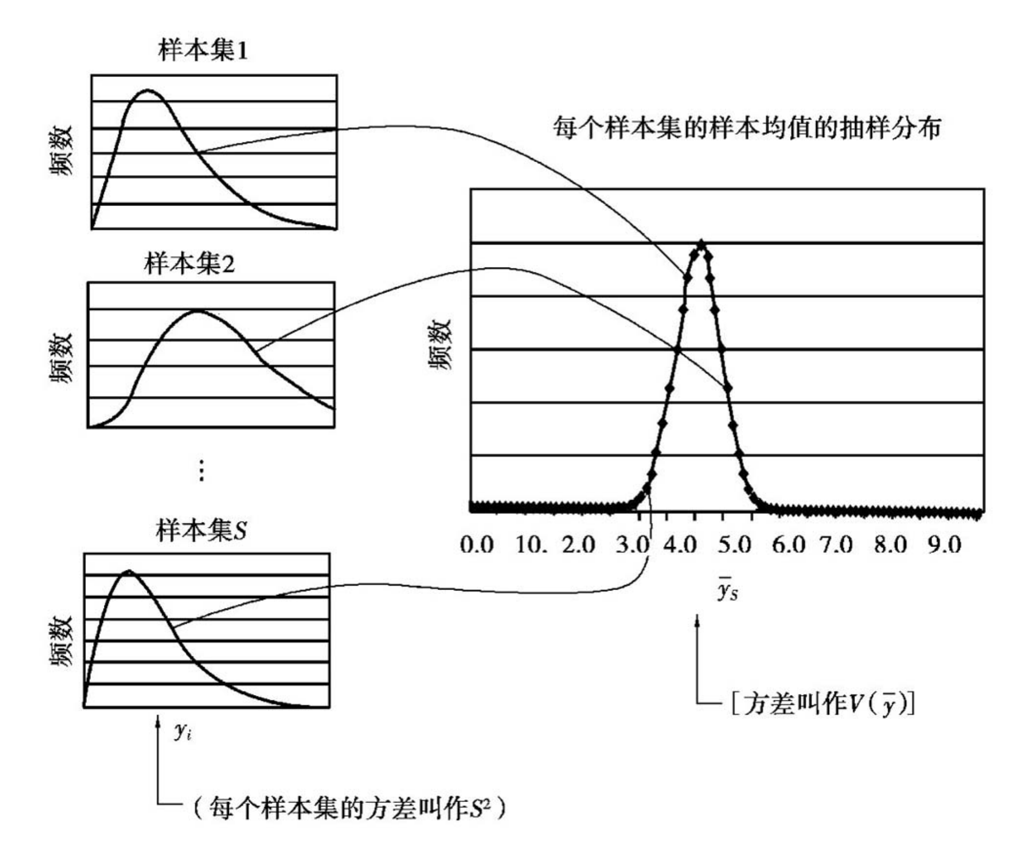


图4.2　样本集中变量y 的分布以及均值的抽样分布

图4.2左边的，就是不同的样本集 （realizations）。一个样本集，就是运用特定的抽样设计选出来的一组总体要素。每一种抽样设计，可以生成许多样本集。可能的样本集数量涉及样本量、框要素的量，以及抽样设计。（不过，请记住，在调查研究中，我们只会用到一个样本集。）

用图4.2，我们还可以描述另一些抽样理论概念。在某种情形下，你可以把每一次调查看作许多可能的概率抽样设计之一的一个样本集，且被用来描述用小写字母表示的样本集的属性。如此，每个样本集都有变量y 的均值与方差。样本均值用$\bar{y}$ 表示，用于描述样本要素中y 的集中度，样本各种值y_i 分布的方差用s^2 表示。在这个抽样设计中，一个样本集的$\bar{y}$ 就被用来估计框总体的$\bar{Y}$ 。同样，一个样本

集的s^2（样本要素方差， the sample element variance）用来估计框总体的S^2。

最后，我们还要用一个样本集的参数估计样本均值（请看图4.3的V（$\bar{y}$））的抽样分布方差。样本均值的抽样分布方差的另一个术语叫作均值的抽样方差（sampling variance of the mean），开方后的值就叫作均值的标准误（standard error of the mean）。如果这些概念对你而言比较陌生，那么，最好现在就花点心思记下来，这有助于你理解后面的内容。图4.3也许有助于你记忆。

	Y变量的分布类型		
	样本集内的分布	框总体内的分布	均值的抽样分布
状态	一个样本集已知	未知	未知
要素量	$i=1,2,\cdots,n$	$i=1,2,\cdots,N$	$s=1,2,\cdots,S$
一个要素的值	y_i	Y_i	$\bar{y}_S$
均值	$\bar{y}$	$\bar{Y}$	$E(\bar{y}_S)$
分布的方差	s^2	S^2	$V(\bar{y})$
分布的标准误	s	S	$Se(\bar{y})$

图4.3　样本集、框总体以及样本均值的抽样分布等关键术语

如果分析者希望用给定调查样本（一个样本集）的均值来估计框总体的均值，就会用到标准误来设定置信值（confidence limits）。置信值用来描述一个置信水平，即所有样本集的均值与框总体均值的差值都会小于某个值（不包括覆盖性、无应答和所有其他误差产生的影响）。例如，假设在SOC中我们估计成人样本的平均年龄为$\bar{y}=42$岁，且该估计的标准误为2.0。如此，置信度就为实际均值加减标准误。例如，如果为年龄均值取68%的置信度，基于标准误，则两

边的置信值为（42-2.0，42+2.0）=（40，44）。其含义为，用相同的抽样方法从相同的总体中抽取样本，在68%的置信度，能够估计实际总体的均值。一般而言，我们采用95%的置信度，根据正态分布，其标准误为1.96，则置信值为（42-2.0×1.96，42+2.0×1.96）=（38，46）。即在95%的情况下，样本与总体的情况是一样的。

样本统计量的标准误用以测量同一个抽样设计下可能样本集统计量的分布与变异性。这类标准误用$se(\bar{y})$表示，即抽样方差$v(\bar{y})$的平方根。即

$$se(\bar{y}) = \sqrt{v(\bar{y})}$$

提示

千万别混淆了不同的方差。每一个框总体，都有自己的Y值分布。S^2是总体要素方差，用给定的样本集的样本要素的方差s^2来估计。样本均值的抽样方差，即$V(\bar{y})$是用样本集的数据即$v(\bar{y})$来估计的。大写字母用于表示来自总体的参数，小写字母用于表示来自样本的参数。

抽样设计或抽样方法会直接影响到标准误和抽样方差。如果给定抽样方法，样本量越大，均值方差就会越小。而整群抽样常常会有较大的标准误。将样本分为不同的组，从组中独立地抽取样本会得到较小的标准误。标准误越大，意味着置信值的区间越宽，反之亦然。

在调查抽样中，对抽样变异性而言，还有一个测量，即抽样比例$f=n/N$，指的是样本量（n）占框总体要素（N）的比率。如果要用样本数据来推断框总体，则需要反转抽样比例来看抽样操作，并预估样本量。$F=(1/f)=(N/n)$。

简而言之，调查的目的在于揭示框总体中变量的未知分布。已知的概率样本（probability samples），意味着框总体要素被选择的概率为非零。如此，我们就可以用一个样本集在给定置信度的条件下来估计框总体的特征。如果我们希望估计均值，则可以用样本集的均值来估计均值的标准误，以及给定置信度下的置信值。

正如我们在第2.3.5节讨论过的，抽样所带来的误差的程度是4个设计因素的函数：

1）样本量。

2）框总体要素入选样本概率的差异。

3）每个要素是单个的、独立的入选，还是以组的形式入选（单个样本还是整群样本）。

4）样本是否用来代表子总体（分层样本）。

本章将围绕这些因素的讨论展开，为展现概率抽样的基础，我们将从最简单的样本设计开始。

4.3　简单随机抽样

简单随机抽样（simple random sampling，SRS），是抽样的基础方法，也是用于比较其他抽样方法方差的对象。简单随机样本意味着框总体的单个要素、成对要素、成组要素的备选概率是一样的。也就是说，首先我们要识别框总体中在样本量为n的情况下每个可能的样本，然后随机抽取。抽样设计将等概率赋予框总体的每个要素，即等概率抽选方法（equal probability selection method，epsem）。

在实践中，没有人会记下抽取样本量为n的所有可能样本。在大样本条件下，如此做实在太耗费时间。例如，在SOC中，从2.28亿美国成人中写下样本量为300的可能样本，将是一个天文数字。在SRS中，可以在要素列表中直接使用随机数字表。SRS常常运用1到N的列表。在1到N的列表中随机选取一定的量，代表总体中的相应要素。如果遇到总体中某个要素被选择到超过1次，我们仍将其入选机会记为1次，并将继续选择，直到选够了n个样本。这就是非替换抽样（sampling without replacement）。

对SRS而言，我们计算变量y的均值为

$$\bar{y} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} y_i$$

即用所有样本值的和除以样本数。某个样本集的样本均值的方差则为

$$v(\bar{y}) = \frac{1-f}{n} s^2$$

评论

一个比例值的估计抽样方差就是一个比例pq/n 的抽样方差的一般估计，这里，$q=(1-p)$，并忽略了有限总体修正。

用样本集中要素y_i 的方差除以样本量，乘以（1-f）。这里的（1-f）是有限总体修正（finite population correction，fpc）。有限总体修正因子是框总体中没有被选作样本的比例或（1-f），这里f 是样本比例。由于采用了非替换抽样，因此会减小抽样方差。如果样本占总体的比例高，则f 会接近于1，如此fpc 就会减小抽样方差。常见的情况是，相对于框总体而言，样本量实在是太小，以至于fpc 可以被忽略。

如果f 很小，则fpc 会接近于1，进而

$$v(\bar{y}) = \frac{s^2}{n}$$

即用对要素y_i 的方差除以样本量。抽样方差、均值的标准误以及置信区间的宽度，取决于两个因素：多大的样本量以及y_i 的变异。如果变异y_i 较大（即s^2 大），那么均值的抽样方差就会大，置信值就会宽。如果样本量增大，就会减少抽样方差，改善估计的质量。

最后，值得注意的是，$v(\bar{y})$ 仅仅是基于样本对样本均值的抽样方差的估计。还有真实总体的值$V(\bar{y})$ 是要估计的，即出自于SRS的$v(\bar{y})$

是$V(\bar{y})$ 的无偏估计。如果置信度为95%，总体均值$\bar{Y}$ 就是$\bar{y}\pm1.96[se(\bar{y})]$ ，用均值加减标准误的两倍。

对比例值而言，有一个类似的公式。在SOC中，假设y_i是成年人对当前经济是否比去年同期更好的评价。如此，是一个y_i二分变量值，1代表更好，0代表非更好。那么$\bar{y}$ 就是一个比例，常记作p，其值始终在0.0到1.0之间。

对这种类型的变量SRS均值的抽样方差为

$$v(p)=\frac{1-f}{n-1}p(1-p)$$

这里，p 的抽样方差不仅取决于fpc 和样本量，还取决于p 值本身。估计$v(p)$比较简单，因为不需要所有的y_i值，而只需要比例值。

那么，一个调查到底要抽多少样本呢？用于确定样本量的方法有多种。其中一种就是找到一个样本量，使其置信值不超过某个允许的值。例如，在SOC中，要求在95%置信度下，认为经济更好的置信值为0.28至0.32，即4个点的区间。那么，这意味着多大的标准误呢？

假设预期比例是0.30，即在置信区间的一半宽度或$\bar{y}\pm1.96[se(\bar{y})]$的一半必须等于0.01，也就是说标准误为0.01。由于均值的标准误等于$\sqrt{(1-f)s^2/n}$ ，如此，我们就知道如何获得样本量n 了。为了计算样本量，我们需要对s^2 的估计值。

我们可以通过之前对相同总体的调查，或针对相同特征的但是另一个稍有不同的总体获得一个近似的s^2，或通过试调查计算一个s^2。我们永远不会有一个精确的S^2，只会有一个近似值，在合理的代价下，我们一定要作出选择。

在比例条件下，问题会更加简单。因为大体上看，$s^2=p(1-p)$尽管会有些误差，由于我们可以预估p，如此就可以预估用于计算样本量的s^2。例如，在BRFSS中，我们可以预估在低收入群体中，肥胖的比例$p=0.3$。那么，$s^2=p(1-p)=0.3(1-0.3)=0.21$。即要素方差为0.21。如果我们希望有95%的置信度，且p的置信区间为（0.28，0.32）（即使用p的标准误为0.01），那么

$$n=\frac{s^2}{(0.01)^2}=\frac{0.21}{(0.01)^2}=2\ 100$$

评论

绝大多数美国的抽样调查都不采用简单随机样本。通常采用分层（第4.5节）和/或整群（第4.4节）样本。

在某些情况下，总体的规模不大，以至于在合理样本量的前提下怀疑fpc不接近于1.0。在这种情况下，就需要调整样本量，以便将fpc纳入考量。例如，要在一个中等规模的镇子，假设$N=10\ 000$成年人，估计肥胖状况。fpc对2 100个样本量有怎样的影响呢？我们也许

会想到，由于总体较小，因而需要的样本量也会较少。如此，调整后的SRS样本量为

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

在本例中，则为

$$n = \frac{2\ 100}{1 + \frac{2\ 100}{10\ 000}} = 1\ 736$$

由于框总体较小，我们只需要1 736而不是2 100就可以达到标准误为0.01的估计。

对样本量，有一种普遍的误解认为，抽取总体的一定比例就行了。其实可以这样看，假设你选择两个总体一定比例的样本，不管其规模多大，满足给定标准误的样本量是不变的，即使其中的一个总体更大。例如，尽管中国的人口是美国的5倍，对相同的估计而言，需要的样本量是一样的。

上述对样本量讨论的前提是给定了关键统计量的标准误。但这并不是通常的情景。通常的情景是，调查者有一笔钱可以做调查。如此，问题就变成了“这笔钱可以用于做多大的样本？”。当遇到这类问题时，在给定统计量S^2的条件下，调查者就要寻找能获得的标准误。

同样重要的是，一项调查通常要询问受访者许多问题，每个调查中都有许多统计量。如果给定样本量，每个统计量的估计精度就会有差异。一个给定的样本量，常常是在可使用经费、最重要的统计量，以及估计精度之间相互妥协的后果。

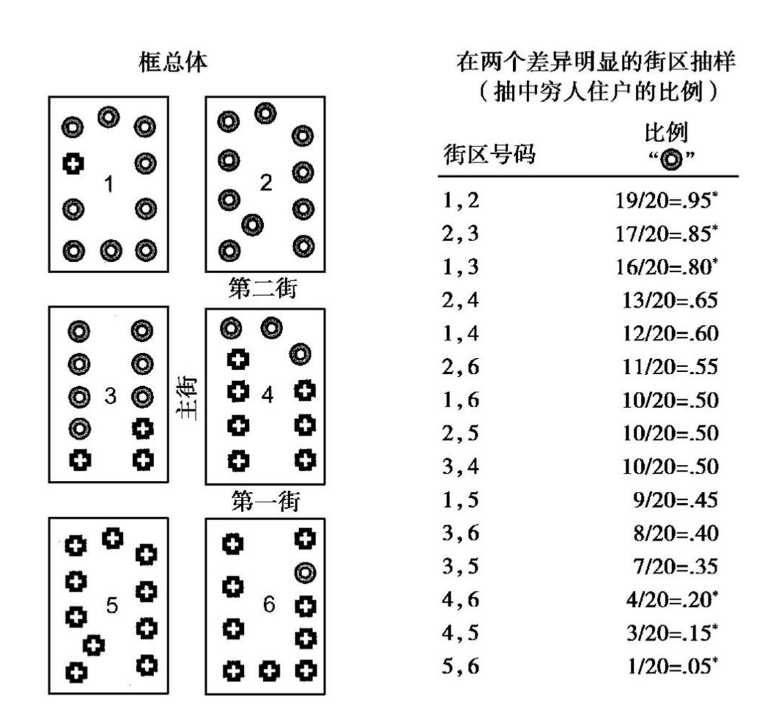
4.4 整群抽样

通常，简单随机抽样的成本很高。整群抽样（cluster sampling）不是直接从框要素中抽样，而是把框要素分成若干组，然后合起来。

例如，即使美国人口登记信息可以列出每一个人，面访调查依然不会使用简单随机抽样方法，因为向全国派遣访员的成本实在太高，尤其是在人口地理分布广泛的情况下。例如，全国刑事犯罪受害者调查（NCVS）中的成人，或者全国教育进展调查（NAEP）中的四年级、八年级、十二年级学生。同样，如果没有总体抽样框的要素列表，制作一个抽样框也会非常昂贵。对单一列表而言，没有分布广泛问题；但对多个列表而言，总会面对这类问题，如全国小学四年级学生的列表；在完全没有列表时也如此，如NCVS中的12岁个体。

用合理的成本建构抽样框的一种方法，就是先抽取群样本，然后对群内的每个要素进行列表。如果是整群抽样调查，就意味着对群内不止一个样本进行了调查，如此也就节省了成本。简言之，在第一种情况下，即总体分布极广，就可以采用整群抽样方法。此外，每个群的规模相等，尽管在现实中不可能，但这样的假设有利于简化均值标准误的计算。

图4.4就是一个简单的整群抽样框。假如一个城市有6个街区居住区，每个街区有10幢房子，总共60幢房子。其中有一半住户是富人，用十字表示；另一半是穷人，用圆圈表示。请注意，穷人总是希望住在穷人的旁边，富人总是希望住在富人的旁边。这是世上典型的居住格局模式。



在两个差异明显的街区抽样
（抽中穷人住户的比例）

街区号码	比例 “◎”
1，2	19/20=.95*
2，3	17/20=.85*
1，3	16/20=.80*
2，4	13/20=.65
1，4	12/20=.60
2，6	11/20=.55
1，6	10/20=.50
2，5	10/20=.50
3，4	10/20=.50
1，5	9/20=.45
3，6	8/20=.40
3，5	7/20=.35
4，6	4/20=.20*
4，5	3/20=.15*
5，6	1/20=.05*

图4.4　一个有着30户穷人和30户富人的居住区的鸟瞰图，其中，十字代表富人，圈代表穷人；在这个居住区随机选择两个街区

假设我们要从这个居住区抽取20户的整群样本，每一个群实际只有2个街区。如果我们从1到6个街区随机抽取2个街区，所有可能的整群样本如图4.4的右边所列。请记住，抽中穷人住户（用圈表示）的比例是0.5。在15组样本中，其平均的比例就是0.5，也就是说，用样本产生了一个对总体的无偏估计。需要注意的是15组样本之间较大的变异性。在15组样本中，有6组的比例大于等于0.8，或小于等于0.2；相对于真值0.5而言，这样的估计是很差的，其估计均值的抽样方差也很

高。也就是说，如果群间差异很大，则整群抽样容易产生较大的偏差。

如何计算整群抽样的统计量呢？让我们用NAEP做例子。假设我们可以获得全美所有4年级学生的班级列表，即$A=40\ 000$；且每个班的学生规模相等，$B=25$；但是，我们没有包含每一个学生的名单表，即$A\times B=N=40\ 000\times25=100\ 000$。不过，我们知道，如果走进任意一个班级，很容易获得该班25位学生的名单。

评论

用一般的统计软件计算整群抽样的标准差，其值会过低。参阅第10章，学习适用的估计值计算方法。

如此，抽样就很简单。采用简单随机抽样方法抽取班级样本a，走访每个样本班级，获得学生名单。假设$a=8$个班级，则学生样本规模$n=8\times25=200$名学生。需要明确的是，这与直接在框要素中进行简单随机抽样不同。如果是在框要素中进行抽取，则有可能抽到200名各不相同的学生，不过，从8个班级中获得的200名学生则不可能做到。

考虑到把每名学生的资料汇集到班级，相关统计量的计算也有所不同。如此，我们需要一些术语和符号来表示学生和班级。假设我们有200名学生中每一名学生的考试成绩，其中在第α班的第β名学生的成绩为$y_{\alpha\beta}$，则考试成绩的均值为：

$$\bar{y} = \frac{\sum_{a=1}^{a} \sum_{b=1}^{\beta} y_{\alpha\beta}}{\alpha\beta}$$

这里，α =1，2，…，表示班级数；β =1，2，…，表示每一个班级的学生数。均值的这种算法，之前讲过。不同的是，首先要加总每个班的成绩，然后再把所有的班级成绩进行加总，再除以样本规模。

这种均值的方差，也不同于简单随机抽样的方差。因为，抽样中，只有班级是随机抽取的，班级不同，$\bar{y}$ 也不相同。除此以外，其他与简单随机抽样相同。这里，我们将班级当作框要素。

如此，则

$$v(\bar{y}) = \left(\frac{1-f}{a}\right) S_a^2$$

式中，S_a^2 为第a 个班级考试成绩均值的变异性。即

$$S_a^2 = \left(\frac{1}{a-1}\right) \sum_{\alpha=1}^{a} (\bar{y}_\alpha - \bar{y})^2$$

式中，$\bar{y}_\alpha$ 是第 α 个班级考试成绩的均值。在整群抽样时，我们使用群间方差，而不是要素间的方差S^2 。

假设8个四年级的考试成绩如下：370，370，375，375，380，380，390，390，则均值等于378.75。

$$S_a^2=\left(\frac{1}{a-1}\right)\sum_{\alpha=1}^{a}(\bar{y}_\alpha-\bar{y})^2=62.5\text{ 且 }v(\bar{y})=\left(\frac{1-f}{a}\right)S_a^2=7.81$$

与简单随机抽样比较，在相同样本规模下，整群抽样会增加均值的标准误。假设采用简单随机抽样方法获得了样本n =200，且均值相同，即378.75；再假设学生间成绩的方差S^2 =500。则

$$v(\bar{y})=\left(\frac{1-f}{n}\right)S^2=2.50$$

与整群抽样的7.81比较，显然整群抽样的方差要大。事实上，是成倍地放大。

$$d^2=\frac{v(\bar{y})}{v_{srs}(\bar{y})}=\frac{7.81}{2.50}=3.13$$

通过班级再抽取学生，我们得到了比直接抽取学生多过3倍的方差。这也意味着它会增大标准差和降低置信度到77%（$\sqrt{3.13}=1.77$）。

统计量d^2 就是设计效应 （design effect），是一个应用广泛的工具，用于评估给定样本量以及整群抽样的效应。抽样效应，就是某种抽样方法的抽样方差，即$v(\bar{y})$ ，与等量样本随机抽样的方差，即$v_{srs}(\bar{y})$ 之间的比值。调查中，每一个统计值都有其设计效应，同一个调查中的不同统计值，其设计效应也不相同。

如果上述例子中所讲的均值方差增加的情况属实，为什么人们还使用整群抽样，而不是要素抽样呢？原因来自于成本，非整群抽样设计的成本太高。为了对学生进行抽样，首先需要制作抽样框，如此就

需要走访40 000个班级，来搜集每个班级学生的名单。此外，如果抽到的200名学生分布在完全不同的班级，与只有8个班级比较，就需要招募更多的访员。因此，在精确性和成本之间就需要取舍。为了在给定预算的条件下完成调查，采用班级抽样的办法就变得可以接受了。

4.4.1 设计效应和群内同质性

相对于简单随机抽样，整群抽样会增大抽样方差。由于每个班级考试成绩的均值不同，也意味着每个班的内部会有更大的相似性或同质性。班级之间的差异越大，班内的差异可能会越小，即同质性越高。

考虑整群效应的另一个方式，就是询问“在同一个群中，增加一个要素会为我们带来什么新的信息？”。在极端的情况下，某班的学生可能会得到了完全相同的成绩。如此，我们只要1名学生的成绩就够了，无需知道其他24名同学的成绩，因为他们的成绩是一样的。这样就会节省很多钱。

测量整群效应的一种办法，就是计算群内要素之间的以及群间的相关性，即测量一个变量值在群内的自我相关性，以及群间的相关性。群内同质性，用roh表示，即“同质性比”（rate of homogeneity）总是一个正数（大于0）。此外，我们还可以把roh与设计效应（即一种抽样方差与等量样本简单随机抽样方差的比）关联起来，$d^2=1+(b-1)roh$。也就是说，抽样方差的增大受到班内学生成绩同质性的影响，也受到每个班学生规模的影响，即b。在上面的例子中，$b=B=25$，即每个班的所有学生都被入样了。其实，我们

也可以在每个班抽取一定数量的学生入样。下一节，我们将讨论次级抽样。

现在我们知道，随着roh 增大，d^2 也会增大。变量不同，roh 也会不同。群内同质性比越大的变量，其均值的设计效应也越大。表4.1列举了不同国家和地区整群概率样本的roh 值。从表中可以看到，社会经济地位的roh 值比较高，而妇女的生育观念和经验的roh 值比较低。在大多数国家，穷人与富人之间的居住越是隔离，就会有更高的roh 值。这也说明社会经济地位变量的设计效应比生育观念和经验的要大。

表4.1　5个国家的不同变量在区域概率调查条件下妇女生育经验的均值*roh*

变量类型	国家和调查年				
	秘鲁 1969	美国 1970	韩国 1973	马来西亚 1969	中国台湾 1973
社会经济地位	0.126	—	0.081	0.045	0.016
人口	0.024	0.105	0.025	0.010	0.025
生育观	0.094	0.051	0.026	0.017	0.145
生育经历	0.034	0.019	0.009	0.025	0.014

数据来源：Kish，Groves，and Krotki，1976.

子样本规模越大，设计效应也会越大。子样本规模b 会放大同质性对抽样方差的贡献。上述学生成绩的例子中，d^2 就相对较大，因为我们把班级内的所有学生都入样了。如果在每个班只选10名学生，即b =10，则d^2 就会下降。当然，为了保证总样本规模，每个班的样本少了，班级的样本数就会从8增大到20，随即，搜集数据的成本也会上升。子样本数量b 对设计效应的影响，也意味着，小子样本规模均值的设计效应会低于全子样本规模均值的设计效应。

最差的情形是，如果roh =1，b =B =25，此时，d^2 就会达到最大值。相反，如果roh =0，则d^2 就等于1。如果roh 为正，整群抽样就不可能优于直接抽取学生的抽样。

Kish和Frankel（1974）关于设计效应对回归系数影响的讨论

Kish和Frankel（1974）发现，许多抽样设计都会影响到样本均值，进而对统计值造成影响，如相关系数、多元相关系数，以及回归系数。

研究设计：在一个模拟学习中，当前人口调查（Current Population Studies）有3 240个初级抽样单位，45 737个家户，从中重复抽出200～300个整群子样本。CPS是多阶段区域概率样本，子样本设计和基础设计一样，也是多级的，且不同的子样本有6、12或30等不同的层，分别对应的平均样本家户数为170、340和847个。计算多元回归系数时，用了两个多元回归公式，每个公式有8个回归系数。通过探讨这200～300个样本的均值和离散值，可以来估计抽样分布。

研究发现：相关系数与回归系数的设计效应常会大于1.0，但小于样本均值的平均设计效应。多元相关系数比其他统计量的设计效应要大。

研究局限：从大样本调查中重复抽取子样本，可以用于模拟复杂统计量的抽样分布。只是，研究本身不能探讨基线调查的设计如何影响了经验结果。

研究的影响：此研究引起了对设计效应影响分析性统计量的广泛注意，这是在之前的复杂调查数据分析实践中被忽视的。

问题是在给定变量和群的条件下，如何通过估计roh 来理解同质性。估计roh 最简单的方法，就是分解d^2 。以上述的班级抽样为例，我们已经知道d^2 =3.13=1+roh （25名学生-1），则可以由此来估计roh 如下：

$$roh = \frac{d^2 - 1}{b - 1}$$

如此，

$$roh = \frac{3.13 - 1}{25 - 1} = 0.088\ 5$$

也就是说，班内具有一定的同质性（roh 越是接近于1，就表示同质性越低），但这个同质性却被子样本规模b 放大了。

在抽样调查的估计中，roh 值总是一个需要提供的参数，且提供的方式总是一样的。首先用抽样的样本要素计算样本设计的$v(\bar{y})$ ，然后将整群样本当作简单随机样本进行计算，

$$v(\bar{y}) = \left(\frac{1 - f}{n}\right) S^2$$

最后，计算d^2 ，接着再计算

$$roh = \frac{d^2 - 1}{b - 1}$$

那么，在实践中，如何估计roh 呢？假设我们有之前类似调查或类似主题调查且总体大致差不多的roh 。如此，在新的设计下，要估计抽样方差，则首先要计算新设计的设计效应，$d^2_{new} = 1 + (b_{new} - 1)\,roh_{old}$ 。这里b_{new} 是新样本中每个群内的样本要素数，roh_{old} 则是之前调查的roh 值。现在，我们要用这个新产生的设计效应计算新样本的均值。

$$v(\bar{y}) = \left(\frac{1 - f}{n}\right) S^2$$

只是，这里的S^2 是之前调查的方差，而n 则是新设计的样本数。

也可以从另一个角度来看d^2 。采用整群抽样，实际是有精度损失的。假设我们从样本规模考虑精度损失，前面的例子中，样本学生数为200名，我们得到了较大的抽样方差的设计效应d^2 =3.13。不过，我们并没有200名简单随机样本的设计效应。我们有的只是由样本规模小得多的简单随机样本所产生的相同的均值方差，即$v(\bar{y}) = 7.81$ 。在这里有效样本规模 （effective sample size）就是$n_{eff} = \frac{200}{3.13} = 64$ 。有效样本规模就是能够产生与实际设计样本规模相等的抽样方差的简单随机抽样样本规模。

4.4.2　样本群内再次抽样

前面，我们提到过减少样本群内的样本要素数（例如每个样本班的样本人数），可以降低考试成绩均值的抽样方差。这样做，实际也是一种妥协，即试图降低群效应对结果精度的有害影响。在上面的例子中，我们抽取了8个班级，每个班级有25名学生，共200名学生。如此，我们可以在每个班的25人中随机抽取10名学生。如果要保证有200名样本学生，就需要将班级数扩大到20个。当b（即每个班的样本学生数）从25降低到10，则样本班级数a 从8增加到20。

那么，从再次抽样中，我们得到了什么呢？那就是样本均值的抽样效率变小了，样本均值的精确度增加了。新的设计效应d^2 =1+（10-1）（0.088 5）=1.80，而不是在群规模为25时的3.13；与此同时，有效样本规模增加到了$n_{eff}=\frac{200}{1.80}=111$ ，而不是先前的64。尽管样本综合规模没有改变，均值的精度却有所增加。

不过，在样本总量不变的情况下，增加群的规模就会增加调查的总成本。例如，说服样本学校配合调查且让样本学生参加测试就很费劲；相反，多一个样本学校的学生参加测试，倒不是什么大事儿。我们在增加研究成本的同时，也降低了因整群抽样而产生的抽样方差。

简而言之，改变从框要素中抽样的单位，例如，从简单随机抽取到抽取要素群可以降低成本，与之相应增大了均值的抽样方差，或者是为获得相同的抽样方差而增大样本规模。抽样方差的增大，可以用设计效应来测量。整群样本均值的方差，来自于同质性比以及每个群内被选中的样本量。

另一些抽样技术，可能会产生与整群抽样相反的效果。现在我们就讨论其中的一种：分层抽样。

4.5　分层抽样

如果样本对总体的子总体具有很好的代表性，那么，概率抽样设计就可以做得更好。分层抽样就具有这样的特点。针对一个框总体，假设我们有每一个要素的相关信息，如此，就可以依据这些信息进行分组或分“层”（strata）。层是抽样框中互斥的要素组。每一个框要素只能被归入一个层中。在分层抽样中，从每个层中独立抽样，一个层一个层地进行。所有层可以使用相同的抽样方法（如简单随机抽样方法），也可以使用不同的抽样方法（如一些层采用简单随机抽样方法，另一些层采用整群抽样方法），分别抽取样本。

图4.5和4.6展现了*分层抽样*（stratification）的方法。图4.5按照字母顺序列出了当前就业调查（CES）中的20名合格样本雇主名单。假设这是某个州某个工业部门框总体雇主名单。除了雇员名单以外，列表中还包括组变量，低、中、高、极高，即雇员数量的等级。例如，Bradburn公司的雇员规模在分类中属于高，即有很多雇员。Cochran公司拥有更多的雇员，属于极高。在4类（*层*）（stratum，strata）规模组中，每一组包含5个要素，即5名雇主。在CES中，我们要测量工资水平，并预计4组之间有差异。如果我们忽视框总体的规模信息，就可以采用简单随机抽样方法。简单地说，假设使用简单随机抽样方法，n =4。图4.5给出了抽样结果，即记录号为9，13，14，18。请注意样本的规模分布，1个在低组，2个在中组，1个在高组。极高组没有被抽中，即框总体中的20%不在其列。另一种方式，或许抽中了4个都在低组或极高组。运用简单随机抽样方法，很难控制抽中哪个规模组。

记录号	姓名	组
1	Bradburn Corp.	高
2	Cochran Inc.	极高
3	Deming Design	高
4	Fuller & Fuller	中
5	Habermann AG	中
6	Hansen PLC	低
7	Hu Electronics	极高
8	HydeBev	高
9	Kalton Group	中
10	Kish Consulting	低
11	Madow USA	极高
12	M.P.H. Bank	极高
13	Norwood LC	中
14	Rubin Inc.	低
15	Sheatsley Co.	低
16	Steinberg Ltd.	低
17	Sudman Inc.	高
18	WalimanAG	高
19	Wolfe & Enix	极高
20	WXM Ventures	中

简单随机抽样,样本规模 4

→Kalton Group
→Norwood LLC
→Rubin Inc.
→WallmanAG

图4.5　按字母顺序排列的20个框总体要素，以及采用简单随机抽样方法抽取的4个样本

记录号	姓名	组	
2	Cochran Inc.	极高	
7	Hu Electronics	极高	分层随机抽样,样本规模 4
11	Madow USA	极高	
12	M.P.H. Bank	极高	
19	Wolfe & Enix	极高	→Wolfe & Enix
1	Bradburn Corp.	高	→Bradburn Corp.
3	Deming Design	高	
8	HydeBev	高	
17	Sudman Inc.	高	
18	WalimanAG	高	
4	Fuller &Fuller	中	→Fuller & Fuller
5	Habermann AG	中	
9	Kalton Group	中	
13	Norwood LC	中	
20	WXM Venture	中	
6	Hansen PLC	低	
10	Kish Consulting	低	
14	Rubin Inc.	低	→Rubin Inc.
15	Sheatsley Co.	低	
16	Steinberg Ltd.	低	

图4.6　按组排序的20个框总体要素，以及采用分层随机抽样方法每个组抽取1个样本

图4.6用了同一个框总体，只是使用分层抽样方法。首先，我们按组对框总体进行排序（注意，可以将极高组排在最前面，接着排高组，以此类推）。然后，在每组内采用简单随机抽样。为保证样本总规模不变，每一组抽取1个样本，既保证了总样本规模等于4，也保证了每一组的抽样比例为1/5。图4.6展示了这类设计的一种实现方式。请注意，每一组由1个样本代表。实际上，只要是采用分层抽样，在这个例子中，就总是每个组（层）由1个样本代表。仅仅由低组或极高组代表情形是不会出现的。就CES调查的关键变量工资水平而言，组间是有差异的。这样抽样可以获得样本均值的更小标准误。

为估计总体值（如均值或比例），就必须结合分层来考量。对总体均值或比例的计算，可以先分层计算，再合并计算。合并方法，取决于分层方法。

4.5.1 按比例配置层样本

按比例配置层样本就是在每一层抽选样本时按照相同的比例，即等比例方法，或称之为*epsem* 。这就是说，如果我们让$f_h=n_h/N_h$为层h 的抽样比例，那么，按比例配置，与前述一样，就是按照同样的比例在每一层抽选样本。换句话说，对所有层而言，$f_h=f$ 。例如，对层h ，按照一定的比例，可能抽选样本数为n_h ，其在层中的比例为n_h/n ，与这一层对总体的比例N_h/N 一致（这里N_h 就是该层的总要素数）。这里，我们让$W_h=N_h/N$ ，代表每一层占总体的比例。

如此，要计算整个总体，就要合并每一层。对层进行加权的一个方法，就是用层占总体的比例W_h 。假设我们要估计总体均值，且已经计算了每一层的均值$\bar{y}_h$ ，要估计的总体均值就是$\bar{y}_{st}$ ，这里st 用来代表“分层”，其计算如下

$$\bar{y}_{st}=\sum_{h=1}^{H}W_h\bar{y}_h=\text{加权后的层均值和}$$

在这个公式中，每一层的均值占总体均值的比例等于其层规模占总体规模的比例。$\bar{y}_{st}$ 的抽样方差也来自于对每一层的合并，只是更为

复杂。假设有一种特殊的分层抽样，每一层都采用简单随机抽样方法，抽取层h 的样本n_h ，则第h 层均值$\bar{y}_h$ 的抽样方差就可以这样计算简单随机抽样条件下，样本规模为n_h 的抽样方差了。

$$v(\bar{y}_h) = \left(\frac{1 - f_h}{n_h}\right) S_h^2$$

层内要素方差，即$S_h^2{}'S$ ，每一层都需要单独估计，即围绕均值$\bar{y}_h$ 的方差。

$$S_h^2 = \left(\frac{1}{n_h - 1}\right) \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

因此，在分层随机抽样中，我们不是像在简单随机抽样中那样计算一个要素的方差，而是计算每一层的方差。

对$\bar{y}_{st}$ ，我们要合并简单随机抽样的抽样方差。

$$v(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2\left(\frac{1 - f_h}{n_h}\right) S_h^2$$

这里，我们用总体比例W_h 的平方对每一层简单随机抽样的方差加权。

那么，分层抽样的均值与简单随机抽样的均值比较，在同等样本规模下，有多精确呢？如我们在整群抽样中做的一样，我们要比较在同等样本规模条件下，分层抽样（在这个例子中，就是分层简单随机

抽样）的抽样方差与简单随机抽样的抽样方差。如此，则分层随机抽样的设计效应为

$$d^2 = \frac{v(\bar{y}_{st})}{v_{srs}(\bar{y})} = \frac{\sum_{h=1}^{H} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_h^2}{\left(\frac{1-f}{n}\right) S^2}$$

在这种情况下，设计效应可能小于1，或等于1，也可能大于1。设计效应的大小，取决于从每一层抽取的样本数，这个数在样本分配时就已知。

Cochran（1961）论分多少层

Cochran（1961）讨论了对一个分层变量而言，到底要分多少层的问题。例如，如果分2层好于分1层，那么分7层会好于分6层吗？

研究设计：假设每层都有相同的规模，Cochran研究了涉及调查变量y 以及分层变量x 的不同情形；探讨了不同层在x 与y 相关的条件下，对均值的抽样方差所展现的分层效应。Cochran也展示了一些真实的例子。

研究发现：大多数情况下，针对一个分层变量，分6层甚至更少就能看到分层效应。下表显示了对同一个变量而言，增加层数，对降低抽样方差的影响，只是这样的效应表现为边际效应递减。如

果调查变量y以及分层变量x 的相关性较大，则需要分更多的层。真实的数据与模拟结果一致。

分层后均值的设计效应

	x与y之间的相关性		真实数据	
层数	0.99	0.85	大学入学	城市规模
2	0.265	0.458	0.197	0.295
3	0.129	0.358	0.108	0.178
4	0.081	0.323	0.075	0.142
6	0.047	0.298	0.050	0.104
无穷	0.020	0.277		

研究局限：研究使用线性模型描述$x-y$ 之间的关系；实际的情形可能更加复杂。

研究影响：适用于对每个分层变量具有较少层的实践。

当然，对比例的估计，与对均值的估计，过程相同。实际上，公式完全相同，只是写法上有些差异。

$$\rho_{st} = \sum_{h=1}^{H} W_h p_h$$

和

$$v(p_{st}) = \sum_{h=1}^{H} W_h^2 \left(\frac{1 - f_h}{n_h - 1}\right) p_h (1 - p_h)$$

以全国教育进展评估调查中的学校特征与政策调查为例，为了简化起见，可以想象某样本框为城市区域所隔开。假设框总体N =8 000所学校，抽选的分层样本为n =480。表4.2显示，框总体被分为了3个层。此外，还假定按照比例在层中抽样，则每个层的抽样比例是0.06。一旦选中样本，我们就要询问480所学校的校长，秋季学期要提供多少非升学班级或“假期式”班级。我们要计算每一层假期式班级数量的均值，以及要素方差s_h^2 。计算结果参见表4.2。

表4.2　按城乡分3层后的按比例分层简单随机抽样结果

层	N_h 框学校数	W_h 总体比例	n_h 层样本规模	f_h 层抽样比例	y_h 样本层均值	S_h^2 样本层要素方差
城市中心的学校	3 200	0.4	192	0.06	6	5
城市其他区域的学校	4 000	0.5	240	0.06	5	4
农村学校	800	0.1	48	0.06	8	7
总和	8 000	1.0	480	0.06		

尽管在所有层的抽样比例相等，但每一层的样本量却不同，层均值和要素方差也不相同。农村学校的假期式班级数的均值较大。随着假期式班级数量的均值增加，方差也会增大。

假设我们只是简单地计算480所学校未加权的均值，就会得到一个有偏的均值，因为其有可能给农村学校的比重太大。在这种情况下，计算未加权均值的算法如下：

$$\bar{y} = \frac{6 + 5 + 8}{3} = 6.3$$

接着，我们可以用每一层占框总体的比率来计算加权的均值：

$$\bar{y}_{st} = (0.4 \times 6) + (0.5 \times 5) + (0.1 \times 8) = 5.7$$

加权后的均值要略低，这是因为未加权的均值过多地代表了农村学校。

加权估计的一个问题是，现有的统计软件不会自动地使用层均值的权数，而是使用每个样本的权数。因此，需要把层均值的权数转变为样本权数。特别是在我们用不同的视角看分层均值的时候：

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{y}_h$$

这个公式并没有说明依据每个样本来计算加权平均数的方法。不过，简单的代数变换，就可以说明如下：

$$\bar{y}_{st} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi} y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} W_{hi}}$$

这里W_{hi} 就是数据集中的权数变量，对所有要素而言，$W_{hi} = \frac{N_h}{n_h}$ ，即第h 层第i 个样本数据的权重。简而言之，加权平均值就是加权总数与权数总数的比。这也是SPSS，Stata，以及SAS计算加权平均数的方法。注意，每个要素的权数就是其抽样比例的导数。

分层均值$\bar{y}_{st}$ 的抽样方差，就是加权后各层方差的和。如果每层使用的是简单随机抽样，则方差可计算如下：

$$
\begin{aligned}
v(\bar{y}_{st}) &= \sum_{h=1}^{H} W_h^2(\text{第 } h \text{ 层均值的方差}) \\
&= W_1^2\left(\frac{1-f_1}{n_1}\right)S_1^2 + W_2^2\left(\frac{1-f_2}{n_2}\right)S_2^2 + W_3^2\left(\frac{1-f_3}{n_3}\right)S_3^2 \\
&= (0.4)^2\left(\frac{1-0.06}{192}\right)(5) + (0.5)^2\left(\frac{1-0.06}{240}\right)(4) + (0.1)^2\left(\frac{1-0.06}{48}\right)(7) \\
&= 0.009\ 20
\end{aligned}
$$

这就是逐层计算方差，然后进行层合并的方法。用此方法计算，就会发现组内的变异性较低。

估计均值5.7在95%的置信区间内标准误的计算，可首先计算均值的标准误。

$$se(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{0.009\ 20} = 0.096$$

然后，再计算均值的95%的置信区间：

$$\bar{y} \pm z_{1-\alpha/2} \times se(\bar{y}) = 5.7 \pm (1.96)(0.096) \text{ or } (5.5, 5.9)$$

当然，也可以计算样本均值的设计效应。要计算设计效应，需要分层随机样本的抽样方差，以及等量样本简单随机抽样的抽样方差。

如表4.2所示，可以计算，480个样本的s^2 =5.51，则480所学校简单随机抽样的均值的方差为

$$v_{srs}(\bar{y}) = \left(\frac{1-f}{n}\right)s^2 = \left(\frac{1-0.06}{480}\right)(5.51) = 0.010\ 8$$

如此，设计效应为

$$d^2 = \frac{v(\bar{y}_{st})}{v_{srs}(\bar{y})} = \frac{0.009\ 20}{0.010\ 8} = 0.85$$

这就是说，按比例分层随机抽样的抽样方差是等量样本简单随机抽样的抽样方差的85%。如此，就会降低标准误（以及置信区间的宽度），如：

$$100 \times (1 - \sqrt{0.85}) = 8\%$$

只要是按比例抽样，就总能得到这样的结果，这也是分层的效应。如果（1）对框的分层本身有意义，即对研究变量而言，分层能够是组间差异最大；（2）按比例，则分层总是能导致更精确的抽样。在调查中，分层也总是使用与需要测量的变量有关的变量，只要有关系，就能使分层有效。

如果没有框总体的信息，又如何分层呢？幸运的是，在实践中，没有信息的情况非常少见。总会找到与调查变量有关的信息用于对总体要素进行分层。例如，如果用名字列表，即使是只有名字，也能大致获得性别和族群的信息。事实上，无需完美的分层，就能达到提高调查代表性精度的目的。

Neyman（1934）论分层随机抽样

Cochran（1934）讨论目标抽样和随机抽样争论中的一个简单问题。

研究设计：Neyman用了其他研究者的例子，大多数是社会问题，如贫困、生育率等。有人认为，针对不同的组，建不同的抽样框，然后选择对总体有“代表性”的样本。也有人提出，给每个人等概率的被选机会是重要的。

研究发现：Neyman观察到，两种设计都有长处，应该可以合并。通过例子和数学演算，他发现，把总体分成他所说的“层”，对不同研究变量的关键统计量的值有正面影响，可以先于随机抽样之前进行。先分层再随机抽样，置信区间会变窄。还有，对关键变量而言，在层内的变异性较大时，提高抽样比例，对任何给定样本规模的抽样，都会使置信区间变窄。

研究局限：对同一个调查而言，优化配置方案不同，统计量的结果也不相同。

研究影响：这篇文章是一个突破，引发了后来对分层概率抽样的广泛应用。

4.5.2 不按比例配置层样本

除了按比例配置层样本以外，也有其他方法配置样本，使其较之等量的简单随机样本有更小的抽样方差。有一种方法，在任何有问题的时候，都可以降低样本均值的抽样方差，以得到最小的抽样方法。这种方法以其发明者的名字命名，即尼曼（Jerzy Neyman）方法。尼曼方法比按比例配置或等样本规模配置要复杂，但其基本设计原理并不十分复杂。

对于一个分层样本而言，尼曼方法有多种非比例配置。任何一种，比直接进行要素抽样的等量样本的抽样方差都要小。如果要使用尼曼方法，我们不仅要知道每一层的W_h，还需要知道标准离差：

$$S_h = \sqrt{S_h^2}$$

或者与之成比例关系的值。

计算每层的$W_h S_h$，然后加总。尼曼方法的样本规模计算如下：

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}$$

这里，样本规模不与W_h成比例，而是与$W_h S_h$成比例。因此，如果层规模比较大，则配置的样本也会多，这点倒与按比例配置样本一致。但如果层内的要素有较大的变异性，即

$$S_h = \sqrt{S_h^2}$$

较大，则该层的样本量也会较大。换句话说，如果层内要素间的方差大，该层所需要的样本量就会大。与此同时，在总样本量给定的条件下，层内要素间方差小的层，样本量就会小。这一点很有意义。如果总体的某些部分有较大的变异性，就需要较大的样本量，以保障其调查参数更加稳定。

采用尼曼方法可以获得较之简单随机抽样更精确的样本配置，甚至超过比例配置。不过，尼曼方法也有缺陷，例如，缺陷之一就是所产生的样本并不必然是比例样本。如果采用比例样本，层间需要有较大的比例差异，故不一定能找到合适的变量来满足层间差异的需求。尼曼方法每次只能处理一个变量的样本配置。如果调查不仅仅是用于搜集该变量的数据，那么，尼曼方法就很难照顾到其他变量。或许针对第一个变量，其样本配置很好；针对其他变量，则不然；如此，针对其他变量的设计效应就会大于1。

如果采用非比例方法配置层样本却不知道层内方差（S_h^2），会有增加整体标准误的风险。例如，一种幼稚的做法就是，不管层规模的差异，给每个层配置相同规模的样本。如果我们把这个方法用到表4.2中，每一层就会有160所样本学校，样本总量为480所学校，与比例样本的样本总量一致。但如此抽样获得的层均值方差将会达到0.011 1，而不是随机抽样的0.010 8，也不是比例抽样的0.009 2。

在抽样设计中使用非比例方法，还有很多的研究工作要做。*尼曼方法*（Neyman allocation）强调了一项调查的单一目的性，即估计总体均值。但是，一项调查同时要达成的目的很多，如估计分值、百

分数、千分数。因此，在设计中仅仅优化一个统计量，不是调查的目的。在实践中，我们就不能依靠尼曼方法。一些认真的研究（Kish，1988），探讨了如何为了多重目的来改善抽样，只是尚没有可供在实践中应用的工具。此外，尼曼方法还忽视了覆盖性、无应答以及测量误差等在层间的差异性。

除了复杂性，分层与仔细配置样本，在抽样设计和抽样中都非常重要。分层抽样可以达成其他任何抽样方法能够达成的获得代表性样本的目的，这是用单纯的简单随机抽样方法无法达成的。在实践中，我们会用到某种形态的分层抽样，且总会获得比非分层抽样更小的抽样方差。

4.6　系统抽样

系统抽样 （systematic selection）是分层抽样的一种简化形式。系统抽样的基本思路是用总体中的每一个数到k 的要素作为样本。在确定了总体规模、样本规模后，就可以依据样本占总体的比例来计算抽样间隔，即k 。确定了抽样间隔之后，就可以随机选择1～k 中的任意一个数，作为起点，每隔k 个数选择一个要素作为样本。

图4.7就是用来例举分层抽样的图4.6的样本，即框总体为20位雇主。根据其雇员规模，由高到低，排列为4个规模层。在分层抽样中，需要从每一个规模层抽出雇主样本，即每层按1/5的比例抽样。按照同样的思路，系统抽样实际上运用了分层抽样的诸多优点。如1/5的抽样比例转换到系统抽样后，就是抽样间隔，即每隔5个抽取一个样本k =1/f =5。在抽样实践中，首先，随机选取起点，这里，即为1～5。图4.7中，随机起点（RS ）=2。第一个被抽中的样本要素就是在排列中

排第2的要素；第二个被选中的要素就是（2+5）=7，即排在第7的要素；第三个被选中的要素就是（7+5）=12，即排在第12的要素；第四个被选中的要素就是（12+5）=17，即排在第17的要素。如果我们继续抽样，则下一个被抽中的就是排在第22的要素，已经超出了框要素范围。

记录号	姓名	组	系统抽样，RS=2
1	Cochran Inc.	极高	
2	Hu Electronics	极高	→Hu Electronics
3	Madow USA	极高	
4	M.P.H. Bank	极高	
5	Wolfe & Enix	极高	
6	Bradburn Corp.	高	
7	Deming Design	高	→Deming Design
8	HydeBev	高	
9	Sudman Inc.	高	
10	WallmanAG	高	
11	Fuller & Fuller	中	
12	Habermann AG	中	→Habermann AG
13	Kalton Group	中	
14	Norwood LC	中	
15	WXM Venture	中	
16	Hansen PLC	低	
17	Kish Consulting	低	→Kish Consulting
18	Rubin Inc.	低	
19	Sheatsley Co.	低	
20	Steinberg Ltd.	低	

图4.7　按组排序的框总体20以及系统抽样，抽样间隔=5，随机起点=2

图4.7显示了这种方法如何总能在列表中抽选到具有代表性的样本。在这里，每5个要素在原始的分层要素就是一个层。因此，系统抽样实际上获得了按照分层方法所取得的同样的代表性比例。

与简单随机抽样或分层随机抽样比较，系统抽样程序更加简单。当然，系统抽样也有瑕疵。例如，抽样间隔k 并不总是整数。如果k

不是整数，怎么办？在系统抽样中，有几种办法来应对抽样间隔不是整数的情形。

评论

再看看图4.7，用系统抽样方法，实际只有5种抽样实现的可能。如果随机选择的结果是从1开始，则抽中的要素序号为1，6，11，16；如果从2开始，则为2，7，12，17；如果从3开始，则为3，8，13，18；如果从4开始，则为4，9，14，19；如果从5开始，则为5，10，15，20。要完成抽样，只有这5种可能。样本均值的抽样方差就是这5种实现方式变异性的函数。在调查中，选取一种实现方式的一个副作用就是不可能获得抽样方差的无偏估计。一般而言，对系统抽样获得的样本，会采用简单随机抽样或分层抽样的方法进行方差估计。

第一种方法是就近取整。例如：假设我们要在12 500个要素的总体中抽取1 000个样本，则抽样间隔为12.5，在实际操作中，其间隔就可以等于13。实际抽到的样本要么为961或962。如果将间隔取值12，则样本量就会增加到1 041或1 042。到底取12还是13，其代价是样本量多一些还是少一些。

第二种方法是将要素清单看做是一个闭环，在计算抽样间隔以后，向上或向下取整（其实上下无所谓）。在上面的例子中，假设我们取12，从12 500个要素中，找一个随机数作为起点。假设起点位置为12 475，从哪儿开始，每数到第12个就抽取一个样本，即12 475，

12 487，12 499为头3个样本，这时实际已经在列表的尾端了。接下来，就需要从12 500开始继续数，到第11，就是下一个样本，以此类推，直到选满1 000个样本。

第三种方法就是直接使用小数间隔。按照小数间隔 （fractional interval）抽样后，再取整。如果抽样者手头有计算器。这种抽样方法是很简单的。选择一个小数的随机起点。如果抽样间隔是12.5，如此，在0.1和12.5之间，随机选择一个3个数字作为起点，即在001与125之间。选中随机起点后，就在倒数第1～2位数，加上小数点。加上小数点以后的随机数，就是随机起点。用随机起点加上抽样间隔，就是第一个样本的位置，以此类推，直到要素列表的终点。抽取完成后，再回头来取整，如此获得的样本量与设计的样本量完全一致。

例如，假设随机起点是3.4，如此，抽样的样本为：3.4，15.9，28.4，40.9，53.4，以此类推。取整的方法是，去除每一个样本号的小数点以及小数点后面的数字，所获得样本就是序号为3，15，28，40，53等。需要注意的是，如此获得的样本，样本间隔并不总相等。这里的第一个间隔为12；第二个为13，接着又是12。尽管实际操作中会如此相异，其平均数总会是12.5。

系统抽样有时候又被称为隐性分层抽样，其做法实际上与分层比例抽样大致相当。因此，从排序清单上获得的系统样本，与简单随机抽样比较，对每一个调查变量而言，其精度都要高一些。

排序清单的一个重要类型就是用地理信息排序。许多不同类型的单位，其特征都与地理分布有关。例如，假设要调查某个州12 500家公司的平均雇员数，根据公司所在位置，从东南到西北对所有公司进行排序，我们就会发现，靠近大都市的公司较之位于乡村的公司，规

模会更大。如此，按照排序的清单进行系统抽样，就隐含了根据公司雇员规模进行的分层，较之简单随机抽样，其精度就会更高。

4.7　实践中的复杂性

在第4.4节讨论整群抽样时，只考虑了很有限的相同规模的例子。在实践中，很少有每个群的规模都一样的情形。如果每个群的规模不等，也有方法进行抽样。那就是进行多阶段的抽样。在大型调查中，多阶段抽样是非常普遍的方法，对此，需要有充分的理解。为此，我们将举一个两阶段抽样的例子。

假设在9个街区里共有315个居住单元，我们要从中抽出21个居住单元。在9个街区中，每个街区的居住单元数分别为20，100，50，15，18，45，20，35，12，那么从315个中抽取21个的抽样比为f=21/315=1/15。

假设我们第一阶段希望抽出3个街区，即f_{blocks}=3/9=1/3；在次级抽样中，采用随机方法抽取居住单位，即f_{hu}=1/5；那么，如果已经抽取了街区，则抽取居住单元总的抽样概率就是抽取街区的概率和抽取居住单元概率的乘积。考虑到两阶段的抽样，则抽取居住单元的抽样概率为：

$$f=f_{blocks}\times f_{hu}=\left(\frac{1}{3}\right)\left(\frac{1}{5}\right)=\frac{1}{15}$$

可以预期的是，平均而言，样本规模将会是（1/15）×315=21。这就是两阶段抽样 （two-stage sample）设计，第一阶段，选择街区；第

二阶段，在样本街区选择居住单元。多阶段抽样，总是有这样的嵌套特征。先做初级抽样，再在获得的初级抽样样本中进行次级抽样。有些调查，例如全国刑事犯罪受害者调查（NCVS）会多达四级抽样，有的更多。

不管怎样，在9个街区抽取居住单元时，很少能准确地抽取所要求的21个居住单元。假设我们抽样到的3个街区的居住单元数分别是100，50，45，则在次级抽样中抽出的居住单元样本将会是：

$$x=\left(\frac{1}{5}\right)(100+50+45)=39$$

即抽出的居住单元样本远远大于21个。如果抽中的街区样本为第4，5，9街区，则样本规模将会是：

$$x=\left(\frac{1}{5}\right)(15+18+12)=9$$

又远远小于21个。如此变异性，就会对完成调查的流程产生影响，也会造成统计分析无效率。对流程的影响是，我们不知道要聘用多少访员。对统计分析的影响是，由于无法预期最终的样本数，进而导致精度的损失。因此，我们要有办法对样本规模进行控制。

4.7.1　用与规模成比例方法进行两阶段整群抽样设计

在街区居住单元数量不等的条件下，进行两阶段抽样的方法是，用等概率的方法抽取居住单元（或任何类型的要素）并控制样本规模，使之符合要求。其中的一个方法就是，先采用随机方法抽选出3个街区，然后在每个街区抽选7个居住单元。如果这样做，不管怎样抽样，都能够满足样本量的要求。不幸的是，它不能满足等概率要求。对居住单元较少的街区，7个居住单元就过度代表了，因此一旦这样的街区被选中，相较于居住单元较多的街区，其入选概率要大得多。

与规模成比例 （probability proportionate to size，PPS）方法，通过改变第一阶段和第二阶段的抽样概率使得要素在多阶段的被选总概率相等，且样本规模不改变，就解决了这样的难题。

表4.3列出了每个街区的居住单元数，累计居住单元数。假设我希望每个居住单元的备选总概率相等，即1/15；街区数量相等，即3。首先，在1～315中随机抽选3个数，假设选中了039，144，249；在累计居住单元数列中，看这三个数落在哪里。我们发现，这三个数分别落在了第2个街区（即039）、第3个街区（即144），以及第7个街区（即249）。这就是被选中的3个街区。

表4.3　9个街区的居住单元列表以及累计居住单元数

街区	该街区的居住单元数	累计居住单元数	被选中的居住单元
1	20	20	001-020
2	100	120	021-120←039
3	50	170	121-170←144
4	15	185	171-185
5	18	203	186-203
6	45	248	204-248
7	20	268	249-268←249
8	35	303	269-303
9	12	315	304-315

现在，这三个街区被选中的概率是不等的。例如，第一个被选中的街区有100个居住单元，相对于第一个街区只有20个居住单元而言，其被选中的概率就要大。就三个被选中的街区来看，在要素层面，第一个被选中的概率为300/315，即3×100/315=300/315，同样，第二个被选中的概率为（设计样本量）×（被选中一次的概率），即3×50/315=150/315。第三个被选中概率要小得多，为3×20/315=60/315。为了使要素被选中的总概率相等，在第二阶段抽选居住单元时，就要使用其第一次抽选街区时期被选中的概率的反概率。以第一个被选中的街区为例，其在第一阶段被选中的概率为

$$f_{block} = 3 \times \frac{100}{315} = \frac{300}{315}$$

为了使得总的备选概率为f =1/15，就要“平衡”其抽选居住单元的概率。在这里，如果要在样本街区内抽选居住单元，则

$$f_{hu} = \frac{1/15}{300/315} = \frac{21}{300} = \frac{7}{100}$$

我们把两阶段放在一起，就知道达成了总概率目标

$$f = f_{block} \times f_{hu} = \left(\frac{300}{315}\right)\left(\frac{7}{100}\right) = \frac{1}{15}$$

也就是说，一旦完成了第一阶段，即选中了街区，就知道了第二阶段的备选概率，即7/100，即7%。同理，可以看到第二个样本街区的备选概率是7/50，其总备选概率为

$$f = f_{block} \times f_{hu} = \left(\frac{150}{315}\right)\left(\frac{7}{50}\right) = \frac{1}{15}$$

即在50个居住单元中，正好要选7个。第三个街区样本的算法也一样。

这类PPS过程产生的是在要素层面的等概率，其结果是，每个群需要被抽样到的样本数，是一样的，进而达成了等概率方法（Equal Probability of Selection Method，*epsem* ）的目标与样本量一致的目标。

在实践中使用这种方法，我们会遇到很多问题。首先，第一阶段抽取随机数字可能会有两个数字落在同一个街区，例如，随机数有可能抽到039，069，110，如此，3个数全部都落在了第二个街区。这种“替换性”抽取街区的方法是可以接受的，但人们更愿意接受抽取3个不同街区的结果。当然，也有不同的方法来达成这个目标，并非所有的方法都在要素层面满足*epsem* 。我们可以采用PPS的系统抽样，只是已经超出了这里的范围（参见Kish，1965）。

其次，是有些样本街区可能没有足够的要素供抽取设计的样本数。假设被抽中的第三个街区只有6个居住单元，那么就不可能抽到需要抽到的7个样本，至少也满足不了简单随机原则。在这种情况下，需要事先有安排，把这类不希望出现的情况清理出来。一个简单的办法，就是合并街区，使其至少有足够的居住单元数，供抽选出7个居住单元来。在上面的例子中，第七街区如太小，就可以与第六街区合并，将总的街区数从9减小到8。如此，就可以在8个街区中运用PPS方法计算累计规模，并按照上述方法进行抽样。

最后，我们也可能遇到过大的街区。例如，假设第二个街区的居住单元数是120而不是100。那么，其在第一阶段的入选概率将会大于1.0，即（3×120/315=360/315）。换句话说，不管哪里是随机起点，它总会被选中。遇到这种情况的一个解决方案就是把其从街区列表中直接剔除，当作必选街区，直接从中选择居住单元。如此，在第二个街区直接使用总体抽样比例即1/15，或由该比例产生的样本数。

$$x_2 = \left(\frac{1}{15}\right)(120) = 8$$

即8个居住单元。此外，剩下的8个街区依然按照PPS方法进行抽样，即选择2个街区，然后在样本街区内进行居住单元抽样，使其总的备选概率为1/15。

4.7.2　多阶段及其他复杂设计

许多调查在基于概率方法设计抽样的基础上结合了上述不同的技术。他们使用分层整群抽样来抽取群样本，甚至子群样本。有时候，甚至会有三或四阶段抽样。他们也可能使用不等概率方法以及其他一些复杂的方法，以使抽样工作在给定资源的条件下便于操作。

在完全没有总体要素清单的情况下，进行多阶段的次级抽样是常见的方法。正如第4.4节讨论的，它是一种经济的做法，即先抽取群样本，然后获取或制作群内要素的清单。不过，如果第一阶段抽到的群样本内的要素数量过大，制作清单仍然会有困难。如此，就需要在初级抽样的基础上进行次级整群抽样，以减少群内要素的规模。同样有可能遇到的情况是第二阶段抽样之后，群内要素的规模依然过大，且

无法制作要素清单，此时就需要进行第三阶段的整群抽样。采用多阶段的方法，直到可以用合适的成本来制作要素清单。

在被称为且使用广泛的“区域抽样 ”（area sampling）中，常会遇到这样的情况。区域抽样，就是抽选区域，用如县、街区、可计数的区域，或者是其他政府定义的地理单元，作为抽样单位。

对多阶段抽样，还需要更多的研究来使其更有效率。抽样理论为我们提供了一些指导，帮助确定第一阶段、第二阶段、第三阶段等的样本数，以达成不同的精度目标。只是，要优化的话，既需要知道主要变量在不同抽样阶段的变异性，也需要知道不同阶段需要付出的成本。在复杂调查的设计阶段（例如，假设可以在第一阶段的区域招募访员，让其到第二阶段的区域去调查，还有联系的次数等，这些都是成本）需要认真研究成本。运用计算机辅助调查，给了调查研究者更多的管理数据，对设计阶段的决策很有帮助，在未来的研究中也会用到。

如果调查目的是既要描述群总体，又要描述群内的要素总体，那么，在群的设计上，就要做更多的研究。例如，参与全国教育进展评估（NAEP）的学校，以及样本学校的学生。搜集到的数据应该可以研究学校，也可以研究学生。近些年，对基于多层次现象的统计分析，有了一些新的重要发展。例如，学生成绩的影响因素，既有来自学生的，也有家长的、教师的、学校的影响。在这种情况下，抽样设计如何能使得测量更加精确，则需要更多的研究。

4.7.3 如何描述多阶段抽样设计：NCVS的抽样设计

我们可以把上面学到的知识综合起来，具体讨论一个调查的抽样设计，即全国刑事犯罪受害者调查（NCVS）。让我们回顾一下，NCVS调查的目标总体是12岁及以上家户平民。我们知道，在美国，我们没有这类人口的清单，所以，抽样设计的第一个问题就是抽样框的制作。设计基于多阶段概率设计。

下面就是NCVS给用户提供的抽样设计说明（ICPSR，2001）。下面我们就转录这个说明，并在每段后面给出评论。

> NCVS的样本来自于多阶段整群抽样，包括了大约50 000个家户样本。当前，每月的样本量大约为51 000个家户。注意，这里是指核实前的样本地址，即包括了可能的空户、非住宅地址。核实过的有效地址是42 000户。初级抽样单位（PSU）为县、合并的县、大都市区。在PSU的基础上进行分层。大PSU自己为层，自动纳入。这类PSU为自代表（SR）样本，因为他们都被选中了。其他的PSU为非自代（NSR）表样本，因为会在其内进行二次抽选后再合并为层，合并的方法是依据人口普查的信息，把具有相似地理和人口特征的区域进行合并。

评论：第一阶段的抽样单位为区域单位，即县，合并的县。有些很大（类似于第4.7.1节所讨论的），以至于全部入选，且被称为自代表单位（如纽约、洛杉矶、芝加哥和其他一些大县）。

> 1995年以前，抽样框基于1980年的人口普查数据。1995年的1—12月，以及1997年，样本逐步采用了1990年的人口普查数据。

从1998年1月开始，整个NCVS的样本都来自于1990年的人口普查数据。

评论：这意味着依据1990年的人口普查数据，采用概率抽样方法获得了初级抽样单位。理想的做法是采用当前数据进行概率抽样。如果数据过时，那么不同层次的抽样单位都会出现样本规模的变异性。

当前的设计包括84个自代表（SR）PSU和153个非自代表（NSR）层，采用PPS方法，每层一个PSU。

评论：在抽样之前，PSU已经被分层。设计中，依据区域、PSU的规模以及其他变量，有153个层。

某个PSU内的居住单元样本，通过两阶段抽取。每个阶段，都保证居住单元为自加权的概率样本，也就是说，在每个选中的区域，在进行任何加权调整之前，每个样本居住单元的总备选概率都是相等。第一阶段是从目标PSU中选择片区（Enumeration Districts，EDs）。片区，是人口普查的时候设立的地理区域，其大小从一个城市街区到几百平方千米不等。通常，区域内的人口数在750～1 500人。片区为系统抽选，与其1980年或1990年的人口规模成比例。第二阶段，将每个选中的片区划分为区块（segments），即每4个居住单元为一块，作为抽样单位。

评论：阶段嵌套抽样是：（1）PSUs（县或合并县）；（2）片区（大约1 000人的普查单位）；（3）区块（大约由4个居住单元组成的

组）。然后，被抽中的块内的每个居住单元都是样本，都要接受访问。

> 区块由1980年和1990年人口普查资料的地址清单组成。不过，如果有普查后新建的居住单元，也要列入。抽取获得许可新建的居住单元，没有获得许可的小片区块，也要抽样。这样的补充，尽管产生的样本数量不大，但却能够使在普查后新建的居住单元也能被代表。此外，一些特殊的居所也要采用特定的程序抽样，如边界的居住单元、宿舍，用它们做一个小的样本。由此产生的样本总量包括50 000个居住单元和其他住所。

评论：用上一次人口普查的地址抽样会有覆盖性问题。上次普查后，也许有人盖了新房子。因此，依据人口普查资料，运用了多个抽样框。因此，NCVS的抽样是一个多框设计（参见第3.7.3节 ）。对其他调查机构而言，他们拿不到上次的人口普查资料，就不得不让访员到街区去，列出样本街区的居住地址列表或居住单元列表，再把这些信息反馈到调查机构，然后才能抽取最后要调查的居住单元和个人。

> 鉴于NCVS调查的连续性，为了避免方位相同的居住单元，需要进行轮换。我们把样本居住单元分为6个轮换组（rotation group），在三年半的时间里，每一组每6个月轮到一次。在由6组组成的轮换组中，实际上构成了6个跟踪调查小组，在6个月的时间里，每个月访问不同的小组。

评论：在确定了所有样本后，随机组成6个小组，然后按月访问样本组。这种“轮换组”的方式，是NCVS轮换跟踪调查的部分特性。在任

一给定的月份，都会有人第一次被访到，有人第二次被访到，有人第三次被访到，以此类推。那些第一次被访到的，就是因轮换进入被跟踪调查的。经过7次访问后，他们又会因为轮换被挤出NCVS样本。

每个访员进入样本居住单元后，首先要建立家户成员列表，然后针对12岁及以上的人，搜集NCVS数据。

如此形成的结果是，个体属于居住单元，居住单元属于区块，区块属于片区，片区属于初级抽样单位。在选择片区之前（空间组），就已经分层了。在片区阶段（同样是空间关系），以及初级抽样单位阶段，都有分层。所有阶段都采用了PPS，使得每一个跟踪组的个体总体上的备选概率是一致的。

4.8　用美国的电话号码进行家户抽样

在美国，大规模的调查都会使用电话调查。如果采用电话调查，最常用的抽样框就是固定电话号码，即假设每个家户有一个固定电话。被抽到的家户，列出其家户成员，然后在家户成员中抽样，访问。由于电话号码样本在美国很普遍，也逐步建设了信息丰富的电话号码抽样框。

在美国，电话号码为10位数，包括：

区号—前缀—后缀

在美国，大约有300个区号里有着正常使用的家户电话号码。前缀通常严格限制在2××到9××的范围，其中×可以是任何的个位数，从0到9。这意味着每个区号下，有800个可能的前缀。对每一个前缀，有10 000个可能的电话号码（即0000—9999）。这个阶段，随时间而迁入迁出的人口，号码变化是最大的。由此我们知道，在这个抽样框中，可能有24亿个可能的电话号码。在2000年，美国大约有1.1亿个家户，其中，8 900万户每户至少有一个电话号码。假设每个家户有一个且只有一个电话号码，如此，实有的电话号码占可能有的比例，只有4%。这就是说绝大多数可能的电话号码，都不是进行调查要用到的电话号码。如果采用最简单的“随机拨号”抽样的话。用所有有效区号加上前缀，再加上四位数的后缀。如此，我们知道能够找到的有效家户电话号码是一个很小的比例。

如此设计，效率太低，不可能投入运用。当前，如果在抽样框仅限于前缀活跃的号码，就会获得更多的有效住户电话号码，在效率上就会有较大的提升。即使如此，我们还可以做些改进来更多地提升效率，即将前缀划分为小段。

在使用中，我们可以使用一个住户电话的号段。通常会以一个区号内的100个号码为一个号段。例如，734-764-8365就在734-764-8300到734-764-8399号段内。一旦发现一个号段内有住户电话号码，那么这个号段就会有较多的住户电话号码。如此，有经验的随机拨号，例如，行为风险因素监测系统（BRFSS）和消费者调查（SOC），就将调查约定在100个有住户电话号码的号段内。通过这类约束，抽样框就会对住户电话号码覆盖不足，但却大大地增加了拨通住户电话的比例。

除了随机拨号抽样框以外，还有一些替代性的电话号码框，不过这些框也有对家户电话覆盖不足的问题。最直接的一个替代就是美国

各地都有的电话号码簿。有些商业公司将这些电话号码簿搜集整理为电子格式的版本。不过，有相当部分的家户电话不在这个清单上，有些是因为新申请的电话，这样，在电话号码簿印刷时，就不在号码簿上；也有的是因为电话的主人要求把自己的号码不要列在号码簿上。在大都市区，大约少于一半的住户电话号码会被登记的电话号码簿上。因此，用电话号码簿作为抽样框，会有较严重的覆盖性问题。

用有效区号和前缀的所有号码作为抽样框，在所有时候，都不是很有效率的做法。早在20世纪90年代，美国的电话系统有两个变化让其增加了区号和前缀。第一，调制解调器的大量使用占用了电话号码。如此，导致了新增电话号码，有些号码就安装在家户，却不用来进行语音通话。如果拨通了，就会收到电子应答或永远没有人接听。

第二，为减少电话服务的费用，引入了地方性的竞争。20世纪90年代之前，一个公司只允许为一个地区提供电话服务（接入服务）。后来变了，在同一个地区，可以有多家公司提供电话服务。此外，新加入的公司，可以允许用户退掉他们之前的合约，且保留其号码。不过，电话号码系统也是记账性系统。一个公司会用指定区号加前缀用于对其客户记账。如果新公司允许其客户退掉老号码，记账系统就会发生混乱。为了更正记账系统，新公司被允许使用特定区号加前缀用作影子账号。当用户用其退掉合约的号码打电话时，其呼叫就会转入其影子账号。不过，用于影子账号的区号加前缀，其号码增加的速度会快于实际家户电话号码的增加。如此，实际用于家户电话的比例就会下降。这一变化，导致了美国激活的区号加前缀的号码的增加。图4.8显示了1986—2008年之间电话系统的变化。图中的x 轴显示了电话号码簿每100个号码号段的数量。数据来自将所有活跃电话号码的区号加前缀划分为100个电话号码的号段，总计有2 500万个号段。1986

年，每个号段内的有效家户电话数为55，2008年则少于25，有效数急剧下降。

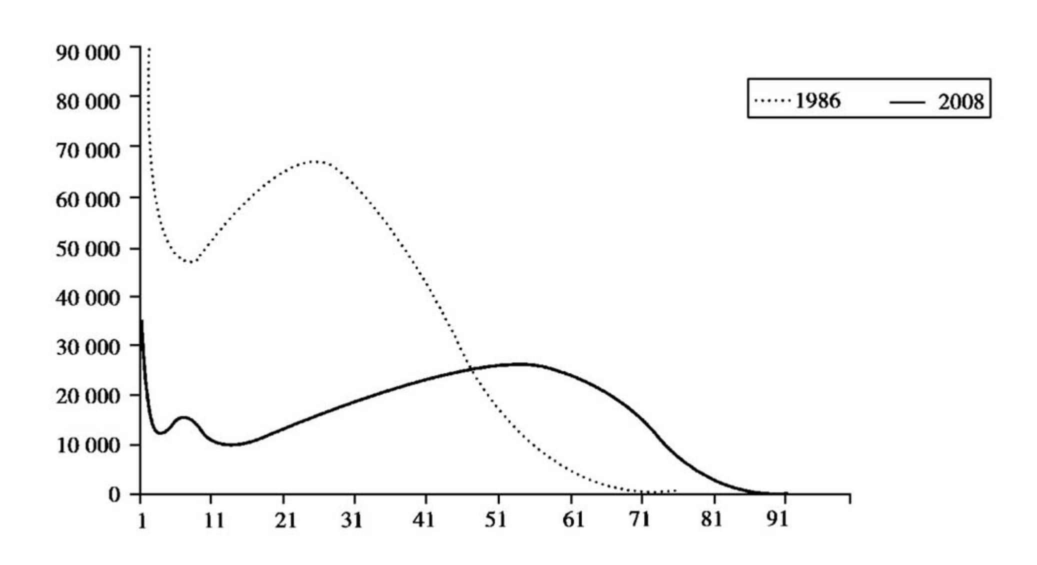


图4.8 100个号段中列在电话号码簿的号码数（1986，2008，数据来源：Survey Sampling Inc.）

简而言之，电话调查面临着选择，要么用一个框（号码清单），其对家户电话的覆盖性不好，但对有效家户的电话号码覆盖比例很好；要么，数字号码框倒是覆盖了所有的号码，不过，有大量的非家户电话号码。在这种情况下，人们创造了多钟替代方法，不过都与分层有关，如过度比例抽样（参见Tucker，Lepkowski，and Pierkarski，2002）、过度整群抽样、双框设计（既有数字号码框，也有电话号码簿清单框，参见Traugott，Groves，and Lepkowski，1987），或其他方法，既运用数字号码框的便利，又试图降低筛选有效家户号码的成本。不过，在美国，进行家户电话调查抽样设计，仍然有需要做的研究。随着手机越来越普及，会有越来越多的问题，如，如何处理固定电话和手机之间的关系，以及如何优化抽样框，如何测量两者之间的重叠度，以及在混合模式下如何尽量减少无应答等，都需要深入认真的研究。

4.9 在家户中抽选个人

在不少家户调查中，要对个体层次进行统计估计。末端抽样常会涉及在一个样本家户中抽选一个人的情形。这一步有许多方法，包括：（a）在家户中建一个抽样框，然后随机选一位；（b）用给定的年龄和性别属性，随机选一位；（c）选一位生日中的月份和日期与调查时的月份和日期接近（或前或后）的成员；（d）按照接受访问的可能性顺序选一位。

Kish（1949）最早提出了一种在样本家户抽选成人的方法，即将家户的所有成员列出，首先是男性，从最年长到最年轻，接着是女性。访员将合格受访者顺序编号（即18岁或以上）。在使用纸版问卷时，对每一个家户，可以使用8种选择方式中的一种。这种做法，使得具有相同合格受访者数量的家户，其备选概率接近，即等概率。也就是说，在同样只有两个家户成员的家户，2位合格受访者的备选概率是一样的，同样，在具有3位家户成员的家户，其备选概率也是一样的，以此类推。不过，在大规模家庭中，8种可选择方式，有可能会使得每位家户成员的备选概率出现差异。这种差异，在计算机辅助访问中，可以使用随机数方法，将被选区间设置为合格成员数，通过随机抽选来得到修正。有人曾经研究过这种方法，当时的主要目的是评估人口普查对家户成员的覆盖问题。研究显示，当家户成员的构成过于复杂时。帮助进行家户成员列表的人，倾向于将与自己关系不太密切的家户成员漏掉（Martin，1999）。还有研究显示，当家户成员中有人曾经违法或者违背当地道德规范时，也会出现同样的漏掉家户成员的情形（Tourangeau，Shapiro，Keamey，and Emst，1997）。这种测量误差，对家户成员列表框会产生影响。

Kish方法需要家庭中有一位知情者，在随机抽样之前，能够帮助识别所有合格的家户成员。Rizzo，Brick，and Park（2004）提出了一种方法，就是用一道问题来进行识别。例如，如果调查仅限于访问成年家户成员，那问题就是“在这里居住的，有几位十八岁或以上的家户成员？”如果只有1位合格的家户成员，那么这位家户成员就会直接被选为受访者。如果有2位，那么就会随机地在两位之中选择1位。如果两位中选择的是“另一位”，访员就会说“另一位被选中了”。如果有多位合格的家户成员，那么软件就会从1到合格家户成员数之间，随机地选择1位。如果1被选中了，访员就会说“你被选中了”。如果选中的是另一个随机数字，访员就会从家户成员的列表中找到与这个数字对应的家户成员，并告诉他被选中了。这样的方法，比较适合于家户成员数为3或更多的家庭使用（在美国，只有15%的家户属于这类家户）。

另一种方法比较适合于家户成员中合格成员数比较少的家户使用。在计算机辅助访问中，这是一种最常用的方法，那就是用随机数字选择来识别性别以及相应年龄的样本个体。举例而言，如果一个家户有2位合格的男性和3位合格的女性。为了抽样，访员首先要询问有几位合格男性在家户居住，有几位合格女性在家户居住。假设有2位男性，3位女性，软件就会随机选择1至5中的一个数。如果选择1，则意味着选择了家户中年龄最长的男性为样本；如果选择2，就意味着选择了年轻的男性作为样本；如果选择3，就意味着选择了年龄最长的女性作为样本；如果选择4，就意味着选择了年龄次长的女性为样本；如果选择5，则意味着选择了最年轻的女性作为样本。如此，这个家户中，每1位合格家庭成员的备选概率就是1/5。一般而言，某个家户成员的备选概率，是家户合格成员的函数。

还有一种方法不会产生概率样本（每一位合格成员的被选机会无法确定）。这种方法，尽量避免使用干扰性问题来建构合格成员框。在对调查做了简单介绍之后，访员会说，“我想访问家里马上要过生日的成年人。”另一种说法是，“我想访问家里刚过生日的成年人。”如果采用这样的方法随机选择受访者，这就给了家户的合格成员同样的备选概率。但如果生日与某个调查变量相关，就会产生偏差。在某些情况下，如对老年的态度，刚刚过生日的人相对于其他人可能会有某种特别的属性。

最后一种方法是在美国的商业调查中采用比较普遍的方法，这种调查通常周期短，且限于回拨的家户。这种方法用来选择受访者中很难找到的人（要么是很难联系上，要么是倾向于拒访）。一个极端的情形是，访员打电话的时候，碰巧受访者在家（Hill，Donelan，and Grankel，1999）。访员希望访问“家里现在在家的18岁或以上的成年家户成员”。如果没有成年的家户成员在家，访员可要求访问“现在在家的18岁或以上的最年长的女性家户成员”。显然，这样的访问仅限于现在在家的家户成员。如此，样本对象就与访员访问的时点有关。如果不是在周末访问，那么那些仅周末才在家的成员就不会被访问到。如果调查中涉及在家的不同模式（如休闲时间利用），则这样的调查就会产生有偏差的结果。

Gaziano（2005）做了一篇方法综述，评估了户内抽样方法，指出了不同技术的关键问题。许多问题都来自假设完全概率方法倾向于提高无应答率议题。一些完全概率方法需要在抽选受访者之前，对知情人提更多的问题。Gaziano注意到，许多研究强调无应答的影响来自访员通过培训所获得的管理访问的能力。因此，无法对替代性方法进行重复性评估。

有鉴于此，不同的调查机构对通过提问题来选择受访者以提高应答率的干扰性议题，有不同的看法。如果要获得概率样本，就要对家户成员有更多的了解。针对这一点，没有争议，争议在于如何选择那个唯一的受访者。在实践中各行其是，主要来自尽量减少选择性偏差以及无应答率，且相信完全概率方法会降低无应答率。

运用概率设计，在一个家户中抽选一个个体。被选中的个体，其备选概率就是家户中合格家户成员的反比例数。这个数值必须要记载下来，用于在要素层的加权，即其在家户成员总数中的比例。如果你要记住这一点，就只需要记住一个个体代表了家户所有其他合格个体。因此，在总体估计时，要把其他个体的权重加在样本个体上，才能让其正确地代表。权重乘以选择家户的权重，再乘以其他的相关的权重（参见第10章 ），就获得了最后的代表性权重。

4.10　小结

在调查方法的研究中，抽样设计是最前沿的领域。大多数的抽样实践都来自概率抽样理论。简单随机抽样是所有其他复杂抽样设计的基础。抽样设计有4个最主要的特征：

1）通过抽选一定量的样本用来估计（其他的类似，抽选的样本量越大，样本方差就会越小）。

2）通过分层抽样，可以整理框总体，将其划分为不同的组，然后再分别抽样。其他事情也一样，分层会降低估计值的抽样方差。

3）通过整群抽样，在抽选群体的同时也获得样本，其他事情也一样。整群抽样，由于同群内的同质性，会增加调查估计值的抽样方差。

4）通过指定变量备选概率，建立不同要素的框。如果备选概率高，那么调查变量的变异性也会大，如此会降低抽样方差；如果不是，就会增加抽样方差。

抽样设计主要受到抽样框的影响。如果没有要素框，最常用的就是进行整群抽样。如果构建抽样框时没有辅助性变量（除要素的识别性变量以外），最常用的就是分层变量。

关键词

区域抽样（area sampling）

整群样本（cluster sample）

置信限度（confidence limits）

设计效应（design effect）

有效样本规模（effective sample size）

等概率抽选方法（epsem）

有限总体修正（finite population correction）

比例间隔（fractional interval）

群内同质性（intracluster homogeneity）

尼曼方法（Neyman allocation）

精度（precision）

与规模成比例概率（probability proportional to size，PPS）

轮换组（rotation group）

样本要素方差（sample element variance）

样本实现（sample realization）

抽样误差（sampling bias）

抽样比（sampling fraction）

抽样方差（sampling variance）

均值的抽样方差（sampling variance of the mean）

无替换抽样（sampling without replacement）

区块（segments）

简单随机样本（simple random sample）

均值的标准误（standard error of the mean）

层（strata）

分层（stratification）

概率样本（probability sample）

随机抽选（random selection）

同质性比（roh）

层（stratum）

系统抽样（systematic selection）

两阶段设计（two-stage design）

进一步阅读资料

Kish, L. (1965), *Survey Sampling* , New York: Wiley.

Lohr, S. (1999), *Sampling* : *Design and Analysis* , Pacific Grove, CA: Duxbury Press.

作业

1．设想有一个小总体，N =8，即8位学生有如下的考试成绩：

$$Y_1 = 72,\ Y_2 = 74,\ Y_3 = 76,\ Y_4 = 77,$$

$$Y_5 = 81,\ Y_6 = 84,\ Y_7 = 85,\ Y_8 = 91$$

（a）计算总体均值：$\bar{Y} = \frac{1}{N}\sum_{i=1}^{8} Y_i$

（b）计算总体方差：$S^2 = \frac{1}{N-1}\sum_{i=1}^{8}(Y_i - \bar{Y})^2$

（c）识别所有抽选样本量为2的可能性（应该有28种可能的样本组合，忽略抽样的顺序），并计算样本均值，画一个样本均值的直方图。

（d）计算由（c）获得的样本的抽样方差，$V(\bar{y}) = \frac{1}{S}\sum_{s=1}^{S}(\bar{y}_S - \bar{Y})^2$，这里$S$是从总体，即$N$=8，抽取样本量为2的总数，$s$角标代表不同的样本。把这个结果与抽样方差$V_{srs}(\bar{y}) = \frac{1-f}{n}S^2$进行比较。

（e）识别所有样本量n=6的简单随机样本组合。计算其均值，画一个直方图。（提示：最简单的方法是运用（c）的结果，看看其与（b）和（c）有什么不同。）

2. 用第一题8位学生的考试成绩，将其进行两组分层：

低分组：Y_{11}=72，Y_{12}=74，Y_{13}=76，Y_{14}=77

高分组：Y_{21}=81，Y_{22}=84，Y_{23}=85，Y_{24}=91

（a）识别所有样本量n=6（每一组的组内样本量n_h=3）的分层简单样本组合。

（b）计算从（a）获得每一个组合的样本均值，画一个直方图。与第1题未分层的样本均值比较，分层后的均值分布又如何？

（c）根据从（b）获得的分层样本均值，计算抽样方差 $V(\bar{y})=\frac{1}{S}\sum_{s=1}^{S}(\bar{y}_S-\bar{Y})^2$，这里，$S$ 为从8位学生总体中抽选样本量 n =6的样本组合数。

（d）将从（c）获得结果，与抽样方差 $V(\bar{y}_{st})=\sum_{h=1}^{2}W_h^2\frac{(1-f_h)}{n_h}$ 比较。（提示：用总体数据计算 S_1^2，S_2^2。）

3. 对不同的抽样设计，统计精度用来评估设计效应。

（a）整群要素样本的设计效应会比1大还是小？

（b）分层要素样本的设计效应会比1大还是小？

（c）在一阶段整群样本中，如果群内所有要素的某个变量具有几乎一样的值，群内的相关系数将接近于什么值？

（d）在一阶段整群样本中，如果一个变量在群内的相关系数为0.016，群规模为10，计算这个关键变量均值的设计效应。

（e）设计效应（d）的值，意味着什么？

4. 假设在一个1 200人的工厂要调查真正因为生病而出现的缺席现象，又假设每年的损失平均天数为4.6，标准差为2.7天，样本来自简单随机抽样。

（a）如果希望损失天数估计值的标准误 $se(\bar{y})=0.15$，样本量的规模要多大？

（b）现在假设按比例分层的层数H =6，由性别、年龄（3个年龄组）构成。如果希望均值的标准误为0.15，按比例分层抽样的样本量会变小，一样，还是变大？简要说明原因。

5. 下表是10个街区的清单，用抽取作为规模量度，抽取PPS系统样本，随机起点为6，抽样间隔为41。

街区	X_α	累计 X_α	抽样
1	32		
2	18		
3	48		
4	15		
5	37		
6	26		
7	12		
8	45		
9	46		
10	21		

6. 一项教育调查，A =2 000所高中，每所学校有B =1 000名学生。等概率抽样的样本量n =3 000名学生，分两阶段抽取。第一阶段a =100所学校，随机抽取；第二阶段，每所学校b =30名学生。在选中的学生中，有30%的说在家可以用电脑。公布的数据显示，这个百分比的标准误为1.4%。忽略有限总体的修正问题，用（n −1）替代n ，估计

（a）样本百分比的设计效应（d^2 ）。

（b）学校内说在家可以使用电脑的百分比的群内同质性（roh ）。

（c）如果样本设计为a =300所学校，每所学校b =10名学生，样本百分比的标准误是多少？（提示：计算新设计的设计效应$d_{new}^2 = 1+(b_{new}-1)\ roh$ ，乘以简单随机抽样的比例方差$\frac{p(1-p)}{n}$ ，这里p 是样本的占比，不是百分比。）

7. 用简单随机抽样方法从一个城市的12 000名选民中抽取了n =10名样本。访问每位样本选民所在的家户，调查其家户是否有中央空调，结果如下。

（a）估计登记选民家户有中央空调的百分比。

（b）估计家户有中央空调的百分比。

样本号	登记选民	是否有中央空调
1	1	有
2	1	无
3	1	有
4	2	有
6	2	无
7	3	无
8	3	无
9	3	有
10	4	有

8. 用分层样本来估计每位大夫所看病人的人数。对总体（N_h）而言，层数规模为（W_h），样本规模为（n_h），抽样比例为（f_h），看病总人数为（y_h），平均每个大夫看病人数为（$\bar{y}_h$），在3层中，每一层的样本要素方差为（S_h^2）：

层	N_h	W_h	N_h	f_h	Y_h	$\bar{y}_h$	S_h^2
年轻	3 200	0.40	192	0.06	1 152	6	5
中年	4 000	0.50	240	0.06	1 200	5	4
老年	800	0.10	48	0.06	384	8	7
合计	8 000	1	480	0.06	2 736		

（a）计算未加权的层均值的均值。

（b）计算加权的分层均值。

（c）计算从（b）得到的均值的抽样方差。

（d）计算（b）得到的均值的95%置信区间。

9. 这里有N =900名病人的医疗记录。用简单随机抽样方法，抽取n =300。其中，210名病人有私人健康保险。

（a）估计有私人健康保险的病人的百分比，计算这个估计值的标准误。

（b）计算总体百分比的95%的置信区间。

（c）这项研究将在另一有N =1 000病人的地方重复。要求对有私人健康保险的样本病人百分比的标准误为2.5个百分点。用简单随机方法抽取样本，需要抽多少样本？为了便于谋划，假设总体百分比为50%。

10. 有一个样本，有a =10群，每群有b =10完访的家户，总个体数和使用手机的人数如下表。

（a）假设样本来自于简单随机抽样，估计手机用户的占比和标准误（假设有效总体修正值接近1。）

（b）由于采用整群抽样搜集数据，故适用于整群样本的方法，估计手机用户p 占比的标准误。

（c）比较从（a）和（b）得到的结果，讨论其差异。

（d）基于整群抽样，计算针对手机用户普遍性的设计效应。

（e）估计roh 。

（f）整群抽样的有效样本规模。

群	总人数	手机用户数	手机用户占比
1	40	10	0.25
2	38	8	0.21
3	25	10	0.40
4	13	5	0.38
5	22	13	0.59
6	28	12	0.43
7	34	10	0.29
8	42	14	0.33
9	20	8	0.40
10	30	10	0.33
合计	292	100	

11. 总体N =10 000个农场的年作物产量方差为S^2 =1 000 000。计算年均产量标准误为$\sqrt{1\ 000}$ 所需的样本量。

12. 一个学院有N =150名教师。院长想做教师薪酬调查，采用系统抽样方法抽取n =20名教师为样本。下面是教师薪酬的列表。

Faculty Salaries (in $1,000)

1	Eng&Prof	m	3	$ 88	51	Eng&Prof	m	3	$ 55	101	Lit&SocSci	m	2	$ 55
2	Medicine	f	3	$ 45	52	Biol&Sci	m	1	$ 49	102	Medicine	m	3	$ 80
3	Medicine	m	3	$ 57	53	Eng&Prof	m	3	$ 57	103	Eng&Prof	m	1	$ 114
4	Medicine	m	1	$ 133	54	Medicine	m	1	$ 118	104	Lit&SocSci	m	1	$ 63
5	Eng&Prof	f	2	$ 71	55	Medicine	m	3	$ 84	105	Medicine	m	1	$ 112
6	Lit&SocSci	m	1	$ 113	56	Eng&Prof	m	3	$ 52	106	Medicine	m	1	$ 93
7	Medicine	f	3	$ 65	57	Medicine	m	3	$ 64	107	Lit&SocSci	m	2	$ 47
8	Biol&Sci	m	3	$ 47	58	Eng&Prof	m	1	$ 75	108	Biol&Sci	m	1	$ 127
9	Lit&SocSci	f	3	$ 39	59	Medicine	f	1	$ 87	109	Eng&Prof	m	2	$ 121
10	Biol&Sci	m	1	$ 74	60	Eng&Prof	m	3	$ 58	110	Medicine	m	3	$ 58
11	Medicine	m	1	$ 88	61	Medicine	f	3	$ 39	111	Biol&Sci	f	3	$ 97
12	Lit&SocSci	m	1	$ 62	62	Medicine	m	3	$ 69	112	Lit&SocSci	m	1	$ 71
13	Lit&SocSci	m	1	$ 49	63	Medicine	f	2	$ 46	113	Eng&Prof	m	1	$ 72
14	Medicine	m	3	$ 88	64	Eng&Prof	f	1	$ 86	114	Lit&SocSci	m	3	$ 29
15	Medicine	m	1	$ 181	65	Medicine	m	3	$ 87	115	Medicine	m	2	$ 167
16	Eng&Prof	m	3	$ 63	66	Medicine	m	3	$ 59	116	Lit&SocSci	m	3	$ 36
17	Medicine	m	2	$ 94	67	Eng&Prof	f	3	$ 44	117	Medicine	m	1	$ 57
18	Eng&Prof	m	1	$ 91	68	Medicine	m	2	$ 123	118	Biol&Sci	m	1	$ 107
19	Medicine	m	1	$ 60	69	Lit&SocSci	f	3	$ 37	119	Medicine	m	2	$ 88
20	Eng&Prof	m	3	$ 55	70	Lit&SocSci	m	1	$ 106	120	Medicine	m	2	$ 87
21	Biol&Sci	m	2	$ 55	71	Lit&SocSci	m	1	$ 91	121	Lit&SocSci	f	2	$ 43
22	Medicine	f	1	$ 106	72	Lit&SocSci	m	1	$ 78	122	Lit&SocSci	m	1	$ 79
23	Medicine	m	1	$ 116	73	Biol&Sci	m	1	$ 77	123	Medicine	m	2	$ 113
24	Medicine	m	3	$ 79	74	Medicine	m	1	$ 90	124	Medicine	m	3	$ 55
25	Lit&SocSci	m	1	$ 61	75	Eng&Prof	m	2	$ 71	125	Medicine	m	3	$ 57
26	Lit&SocSci	f	3	$ 37	76	Medicine	f	3	$ 42	126	Eng&Prof	m	3	$ 56
27	Medicine	m	2	$ 72	77	Medicine	f	2	$ 59	127	Eng&Prof	m	2	$ 65
28	Eng&Prof	m	1	$ 105	78	Eng&Prof	m	2	$ 49	128	Medicine	m	2	$ 42
29	Medicine	m	2	$ 79	79	Biol&Sci	m	1	$ 83	129	Medicine	m	1	$ 102
30	Medicine	m	1	$ 61	80	Lit&SocSci	m	1	$ 34	130	Medicine	f	3	$ 40
31	Medicine	m	1	$ 86	81	Medicine	f	3	$ 42	131	Eng&Prof	m	3	$ 53
32	Biol&Sci	m	1	$ 103	82	Medicine	m	2	$ 97	132	Medicine	m	3	$ 82
33	Lit&SocSci	m	1	$ 48	83	Medicine	m	1	$ 109	133	Medicine	m	2	$ 64
34	Eng&Prof	m	2	$ 64	84	Lit&SocSci	f	2	$ 48	134	Eng&Prof	m	1	$ 72
35	Eng&Prof	m	1	$ 78	85	Medicine	m	1	$ 47	135	Biol&Sci	f	3	$ 36
36	Medicine	f	2	$ 53	86	Eng&Prof	m	2	$ 45	136	Lit&SocSci	f	1	$ 66
37	Biol&Sci	m	1	$ 85	87	Medicine	m	3	$ 83	137	Medicine	f	3	$ 66
38	Eng&Prof	m	1	$ 61	88	Medicine	m	2	$ 51	138	Medicine	m	2	$ 102
39	Medicine	m	1	$ 106	89	Biol&Sci	m	1	$ 78	139	Biol&Sci	m	1	$ 103
40	Lit&SocSci	m	2	$ 60	90	Lit&SocSci	m	1	$ 70	140	Medicine	m	1	$ 148
41	Biol&Sci	f	1	$ 73	91	Eng&Prof	f	2	$ 46	141	Lit&SocSci	f	1	$ 60
42	Medicine	m	1	$ 70	92	Eng&Prof	m	1	$ 85	142	Lit&SocSci	f	3	$ 46
43	Medicine	f	3	$ 32	93	Lit&SocSci	m	1	$ 53	143	Lit&SocSci	f	1	$ 57
44	Lit&SocSci	m	2	$ 49	94	Medicine	f	3	$ 40	144	Medicine	f	2	$ 50
45	Eng&Prof	m	3	$ 43	95	Eng&Prof	m	1	$ 87	145	Lit&SocSci	m	1	$ 90
46	Medicine	m	1	$ 75	96	Lit&SocSci	m	1	$ 71	146	Eng&Prof	m	3	$ 63
47	Lit&SocSci	m	1	$ 92	97	Medicine	m	1	$ 75	147	Eng&Prof	m	1	$ 80
48	Medicine	m	2	$ 107	98	Biol&Sci	m	1	$ 85	148	Medicine	m	3	$ 56
49	Biol&Sci	m	2	$ 57	99	Lit&SocSci	m	2	$ 50	149	Medicine	m	1	$ 72
50	Medicine	m	2	$ 114	100	Medicine	m	3	$ 118	150	Eng&Prof	m	1	$ 96

（a）抽样间隔是多少？

（b）可能的随机起点是哪个数值？

（c）抽出20名教师样本。

（d）用（c）抽出的样本估计教师的平均薪酬。

13. 要去一个集会场所参加机会，所有参会人员必须通过两个入口中的一个才能进入。在两个入口分别采用抽样间隔25和75进行系统抽样，这里，入口2比入口1进入的人数要多3倍。对样本参会者，询问他们从多远的地方来参会。得到的10个访问结果如下：

入口	1	1	1	1	1	2	2	2	2	2
距离	12	34	450	75	240	470	455	24	16	200

（a）估计平均距离及标准误。为了计算标准误，对任何你曾有的距离假设做调整。

（b）对估计的均值，计算其95%的置信区间。

（c）有计划在另一个集会场所进行抽样，那里有3个入口，$S_1=100$，$S_2=200$，$S_3=400$。通过入口1的人数是入口2的2倍，通过入口2的人数是入口3的3倍。即$W_1=0.6$，$W_2=0.3$，$W_3=0.1$。假设我们在入口1的抽样间隔是25，在另外两个入口的抽样间隔应该是多少？

14. 从一个大规模总体中抽取一个两阶段的整群样本，样本量$n=1200$，群数$a=60$，$S^2=500$。在公开的报告中说，$v(\bar{y})=9$ 。一位同事说，均值的方差，计算有误。你同意吗？并用计算结果对你的判断进行解释。（提示：报告中方差的roh 是多少？）

15. 判断下面陈述的对错，并给出你的自己正确陈述。

（a）所有概率样本都是可以测量的。（如果认为是错的，举例说明什么样的样本是不可测量的。）

（b）所有等概率样本都是可测量的。

（c）在同等样本规模的条件下，整群样本总是不如简单随机样本精确。

（d）为提高精确性，分层抽样在于找寻层间的异质性，而整群抽样则是找寻群的同质性。（如果认为是错的，说明分层抽样和整群抽样，都在找什么？）

（e）抽样设计不同，要素方差（S^2）也会不同。

16. 在一些商业机构的100 000名员工中，采用PPS方法抽取1 000名员工样本，其中，每个机构10名员工。商务部2001年对商业机构的员工统计可以作为参考。

（a）如果2001年某机构的员工数为40，那么备选概率是多少？

（b）如果有一批机构2001年的员工数为56，14，84，92，16，8，30，且要从一层中抽取2个机构，如何抽选？

17. 下面是从总体N =270个街区中采用简单随机方法抽选的n =20个样本街区，每个街区的居住单元数如表所示。

i	1	2	3	4	5	6	7	8	9	10
y_i	31	21	2	19	35	0	17	0	27	27
i	11	12	13	14	15	16	17	18	19	20
y_i	1	8	31	11	59	11	5	47	0	54

（a）计算每个街区的平均居住单元数。

$$\bar{y} = \frac{y}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

（b）计算要素方差 $s_y^2 = \left[\sum_{i=1}^{n} y_i^2 - (y^2/n)\right] \Big/ (n-1)$ ，样本方差 $\mathrm{var}(\bar{y}) = (1-f)s^2/n$ ，以及均值标准误 $se(\bar{y}) = \sqrt{\mathrm{var}(\bar{y})}$ 。

（c）估计所有街区的居住单元数 $N\bar{y}$ ，及其标准差 $se(N\bar{y}) = N \times se(\bar{y})$ 。

（d）计算估计均值的95%的置信区间 $\bar{y} \pm t_{(n-1,1-\alpha/2)} \times se(\bar{y})$ 。

（e）如果样本量从n =20增加到n =50，那么估计值的标准误是多少？

（f）如果希望标准误为3.5，那么，需要多大样本量？

18. 假设教师总体N =150，将150名教师编号，从001～150，采用简单随机抽样方法，抽选n =15名。忽略随机号码中的空值，即某号码没有对应的教师。计算如下参数：

（a）样本均值：$\bar{y} = y/n = \sum_{i=1}^{15} y_i \Big/ n$

（b）样本要素方差：$s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ (n-1) = \left(\sum_{i=1}^{n} y_i^2 - (y^2/n)\right) \Big/ (n-1)$

（c）均值的抽样方差：$\mathrm{var}(\bar{y}) = (1-f)s^2/n$

（d）均值的标准误：$se(\bar{y}) = \sqrt{\text{var}(\bar{y})}$

（e）样本均值95%的置信区间：$\bar{y} \pm t_{(1-\alpha/2,n-1)} \times se(\bar{y})$

19. 下表给出了总体的分层信息。我们希望计算有特征的单位在总体中的占比。总样本量n =30。

层h	总体中计数N_h	某特征的占比P_h
1	100	0.9
2	200	0.5
3	300	0.1

（a）计算等比配置的层样本规模（不要把结果求整）。

（b）用（a）的结果计算每一个估计占比的标准误。每一个估计标准误的自由度是多少？（保留小数点后4位，可以忽略有限总体修正因素。）

（c）在所有样本配置方法中，尼曼方法会得到最小标准误，但是调查者不一定用它。说出两个理由，解释为什么我们并不总是使用尼曼方法。

（d）如果你要公开的估计值，你会如何配置样本，解释理由。

5　数据搜集的方法

不管抽样设计有多好，如果后续的调查设计与调查目的不匹配，调查的结果就会产生偏差。调查设计包括编制和测试测量工具，通常指问卷（参见第7章 和第8章 ）；如果使用访员，还包括招聘、培训访员，以及对访员进行督导（参见第9章 ）。本章专注于影响调查的另一个决策，即用什么方法搜集数据。

“搜集数据”在字面上容易产生误解，即容易被认为数据已经存在，只需要将其弄到一起就行了（参见Presser，1990）。调查数据，通常产生于问卷调查或完成问卷调查之后。换句话说，数据是数据搜集过程的结果，而数据搜集过程的重要性又常常被低估。尽管如此，我们还是会运用“数据搜集”这个术语。如图5.1所示，所有的设计，都需要通过数据搜集过程，才会得到调查估计值。第2章已经讨论了数据搜集中两个推论步骤。尽管将数据搜集看做是设计之后的操作步骤，在产生用于分析的数据的过程中，对用经验来检验理论而言，它却是非常重要的因素。这一章将集中讨论数据搜集方法选择中的一些决策，以及这些决策在调查中对成本和偏差的影响。

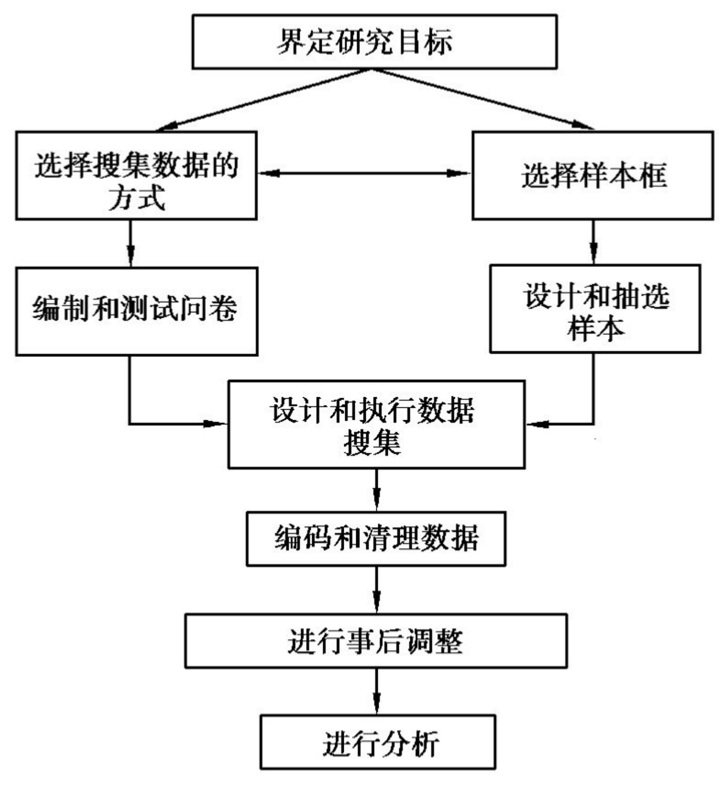


图5.1　调查的过程

传统上，调查依靠三个基础的数据搜集方法：邮寄纸质问卷给受访者，受访者填好问卷后再寄回；由访员通过电话呼叫受访者并进行访问；由访员到受访者家里或办公室进行面对面访问。计算机革命对这些传统的方法要么有替代，要么有改变，并形成了一些混合方法。不仅是方法上的扩展，研究者也能将多种方法结合起来以降低成本或减少误差。如此，也给设计带来了更多的选择以及更多需要做的决策，当然还包括决策依据。本章将关注的正是涉及数据搜集方法的决策以及决策执行的成本与误差。影响决策的，有两个基本因素：

1）对某个具体的研究问题而言，最合适的方法选择是什么？

2）具体的方法对调查误差和成本的影响又是什么？

要回答这些问题，调查设计者和数据分析者都需要理解数据搜集的过程。从设计者角度看，外在的因素常常会影响到数据搜集的方法，如执行成本、数据质量、无应答、覆盖性等。从分析者角度看，知道数据是如何搜集的对评估数据质量至关重要。因此，无论是面对模式选择，还是已有的决策，理解数据搜集方法的细节对理解数据质量都非常重要。

5.1　数据搜集的备选方法

在过去的25年甚至更长时间，调查方法专家们创造了许多搜集数据的新方法。如果考虑到调查有时候会同时使用多种方法，则方法的数量甚至更多。例如，与样本单位的联系方法不一定就是搜集数据的方法。访问不同，也许会用到不同的方法；或者在开始时用的是一种方法，接下来的时候，用的是另一种方法。例如，在过去的时间里，全国药物使用与健康调查（NSDUH）对一些访题采用了面访方法，对另一些访题则采用了自访方法。在历时调查中（对同一受访者多次搜集数据），在第一波调查中使用一种方法，在随后的波次中，可能使用不同的方法。例如全国刑事犯罪受害者调查（NCVS）对受访者要访问7次，初访时采用面访，后面则采用电话访问。第一次采用面访是为了提高应答率，并激励受访者为后面的访问提供确切的信息。随后的访问采用电话访问，是为了节约成本。

对调查机构而言，在最初的几十年里，“模式 ”（mode）意味着要么是面访（或个访），要么是邮寄调查。到1960年代后期，电话调查变得更加普遍了，在后来的几十年里，其普及程度更高。早期的模式比较研究，针对的主要是这三种模式。1970年代讨论的模式效应，

也主要是针对面访与电访而言的。后来才扩展到了其他模式。例如Goves和Kahn（1979）比较了1个全国性面访（74个县和大都市区的整群）样本和2个全国性电访样本（其中1个是同一初级区域的整群样本，1个是采用电话号码的随机拨号样本）。他们比较了应答率、成本、覆盖率，以及测量误差。总体而言，在用电访数据补值的条件下，电访和面访所得到的结果相似。

最近这些年的发展，随着计算机的应用，更多搜集数据的方法得到了发展，调查模式也得到了更宽的拓展，包括多种方法的结合或混合设计。最常见的数据搜集方法如下：

1）计算机辅助个访（computer-assisted personal interview，CAPI），即在计算机屏幕上显示调查问题，访员向受访者念出访题，并记下受访者的应答。

2）计算机辅助语音自访（audio computer-assisted self-interviewing，audio-CASI，ACASI），即受访者自己操作电脑，电脑屏幕上显示访题，计算机语音念出访题，受访者自己记录应答。

3）计算机辅助电话调查（Computer-assisted telephone interviewing，CATI），与CAPI相对应。

4）互动式语音应答（Interactive voice response，IVR），与ACASI相对应，因此，也叫做ACASI，或者T-ACASI，即计算机通过预先的录音扮演在电话中向受访者提问的角色，应答者运用电话上的拨号键盘记录答案。

5）互联网调查 （web survey），即计算机控制的在线调查。

图5.2列出一些当前使用的、更复杂的数据搜集方法，以及方法之间的关系。所有这些方法都可以在传统的邮寄问卷调查、电话调查以及面访调查中找到其根源。看这张图最好的方法是从左到右，随着时间的推移，我们看到计算机发挥着越来越重要的影响。最初的邮寄问卷调查随着文字识别系统 （optical character recognition，OCR）的应用而得到提升，随后，又有了针对手写记录的智能识别系统（intelligent character recognition，ICR）。尤其是在商业调查中，不仅用机器阅读问卷和对应答进行编码，后来还发展到用传真机发送自访纸版问卷。另一些从邮件问卷调查发源的发展如计算机辅助抓取数据。对那些有计算机的人（职场人士）而言，研究者可以给他们寄一个磁盘，磁盘中有调查问题以及记录数据的软件（邮寄磁盘方法） （disk by mail）。受访者完成调查后，再将磁盘寄回。有了互联网以后，问卷就可以通过电子邮件发送给受访者（E-mail method，电邮方法），稍后又有了网页方法 （web method）。这些方法都被称为计算机辅助自访问卷 （computerized self-administered questionnaires，CSAQ）。网页调查又被扩展为了一些更加广泛的方法，包括选择和邀请受访者（参见Couper，2000）以及执行调查的方法（参见Couper，2008b）。

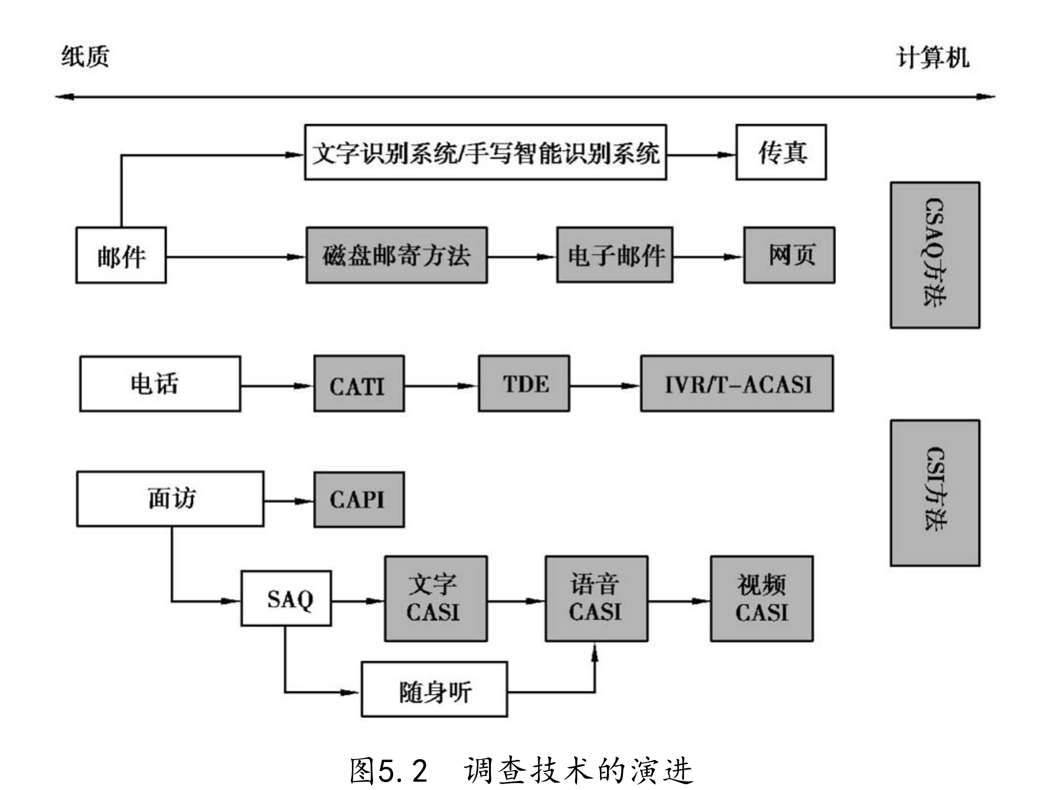


图5.2　调查技术的演进

运用电话网络方法的演进，首先的发展是接入主机的访员电访终端，后来有了接入微机的终端，以及接入个人电脑网络的终端。CATI软件在访问中为访员呈现访题，访员念出访题，软件接收与校验访员键入的应答。一些商业调查对调查数据的要求不高，也会使用电话键盘数据输入（touchtone data entry，TDE）。TDE是邀请受访者拨一个由调查方提供的免费电话，收听访题，并用电话上的键盘输入应答。在这个领域的一个发展就是互动式语音应答（interactive voice response，IVR或T-ACASI）。在这种模式中，访员先给受访者打电话，确认受访者接受访问后，转为访员应答（参见Steiger and Conroy，2008；Tourangeau，Steiger，and Wilson，2001）。采用呼出转接和呼入IVR模式，对调查误差是有影响的，特别是无应答误差。

面访调查的演进有两条路径。第一条是对纸版问卷的替代；另一条是面访中有时会拿一部分出来在访员在场的情景下让受访者自访。

计算机辅助个访（CAPI）用笔记本电脑显示访题和记录应答，替代了纸版问卷。对自访问卷（self-administered questionnaires，SAQ），有时候就会交给应答者去完成，而计算机辅助的方式多种多样。第一种是使用了很短的过渡技术，用索尼的卡带随身听播放访题，用纸记录应答。第二种是各种类型的计算机辅助自访（computer assisted self-interviewing，CASI）方法，包括用笔记本电脑呈现文字访题和记录应答（text-CASI），也有用语音播放访题（ACASI）的，以及作为测量的一部分，用图像呈现（video-CASI）的。

图5.2显示了从较旧的（以纸版为基础的）技术到较新的（以计算机为基础的）技术的进步（关于不同技术的讨论，参见Couper and Nicholls，1998，以及Couper，2008a）。这个列表并没有穷尽所有的方法，新的方法以及既有方法的各种变形在不断出现。显然，要讨论调查中的“模式效应”，积累还不太够。当我们要比较模式的时候，还需要将方法的特征及其与既有观察之间的关系表述清楚。搜集数据的不同方法意味着不同的维度：访员的介入程度，与受访者的互动程度，沟通中涉及受访者隐私的程度，沟通的渠道，以及技术的使用程度。下面，我们将详细讨论这些议题。

5.1.1　访员介入的程度

有些调查设计直接采用了访员与受访对象或知情人之间面对面互动的方式。例如，在面访中，访员向受访对象当面念出访题。另一个调查设计（如邮寄调查）中，根本不采用访员。在这两种极端的情况之间，有许多不同的方法，例如各种自访问卷（SAQs）也可以是由访

员主导调查的一部分。在调查中，访员在场，就像全国药物使用与健康调查（NSDUH），尽管访员要到场，但针对药物使用等敏感性问题，就采用了计算机辅助语音自访（ACASI）。还有一些其他的形式，如在电话访问中，在进入自动访问之前，访员可能会先打电话确定受访对象或说服受访对象接受访问，这就是招募-转换互动式语音应答（IVR或T-ACASI）模式。

访员介入的方式对调查质量和成本都有很大的影响。在访员主导的调查中，需要有接受过培训的、有能力的且有动力的访员，还要有督导员，以及支持人员。在一般情形下，在有访员参与的调查中，访员的费用要占相当的部分。是否使用访员，对整个调查的组织工作有重要影响。

不过，有访员参与，通常能有效地找到样本对象，进而潜在地影响了调查统计的无应答误差。第6.7节展示了通过访员对受访对象强调调查的重要性进而提高受访对象合作程度的例子。有访员回拨电话可以有效地提醒没有完成访问的受访对象。同样，从找到并识别合格样本到家户内的样本选择，访员也是有效执行抽样设计的因素。

访员还有助于向受访对象澄清问题，解释问题，并激励受访者准确地完成访问。这与调查问卷中的访题无应答直接相关。不过，访员在场有时候对应答也会有害，特别是在遇到敏感访题的时候（参见第5.3.5节 ）。在访员主导的调查中，访员的种族、性别所影响的对访题的阐释，以及提问时完成任务的倾向也会影响到受访者的应答（参见第9.2.2节 ）。的确，几乎所有的访员效应都涉及在受访对象面前访员对社会测量重要性的界定。例如，一道关于种族态度的访题或许会因为访员自己的种族不同而有不同的阐释。因此，访员既可能提高也可能降低调查数据的质量。

5.1.2 受访者介入的程度

在面访中，访员与受访者有较高程度的互动。一般情况下，访问会在受访者的家里进行，受访者应答，访员记录应答。电访，如BRFSS，与受访者之间的接触就比较少。另一些调查，如农业的测产调查，与受访者的互动也比较少，访员只需要获得农民的许可，容许其到田间测产就可以了。还有一些调查，特别是那些只需要管理记录的，就完全不需要与样本单位接触，访员只需要从记录中提取所需要的数据就好了。例如，NAEP调查中的部分，只需要定期从样本学校的管理记录系统中提取样本学生的高中课业成绩。2005年有一次课业研究，2009年有另一次。

访员与受访者之间互动数据越多，理论上，就意味着研究者在测量过程中的控制就越多。如果调查仅仅基于管理记录，那么，在调查开始之前，就需要对记录的数据质量进行确认。同样，管理记录对样本总体的覆盖性就不在研究者的掌控之中了。另一方面，与受访者互动产生的数据越多，对特定情景下产生的数据的存疑度就会越高。

如果用到管理记录，就像要了解问卷的制作过程一样，还需要将记录保存系统的相关特征记录下来（Edwards and Cantor，1991），不仅包括受访者理解问题的能力，也包括其对相关记录（忆）的可及性。此外，也不是所有的受访者都有能力拿到所有涉及访题的记录。因此，如果涉及多个受访者，这样的方法用起来就会有困难。要完成调查，还得靠受访者与样本单位内（如家户、商业机构或农场）其他人的关系。

在这方面，研究者可控制的部分很少。测量的特征与管理记录的可用性决定了是否需要与受访者有更多的互动或采用较少的指导。询问农民作物的产量或直接在田间进行客观测量，取决于访题。有一种趋势是，将客观测量（如管理记录、直接观察、交易数据等）与受访者的应答结合以来。人们越来越多地认识到受访者应答与管理记录之间是互补的，进而能弥补信息来源单一所带来的问题。不过，每一种信息来源都有误差问题。

5.1.3　隐私性的程度

调查展开的情境是多样的，部分受搜集数据方式的影响。访问的方式不同，完成调查涉及受访者隐私的程度也不同。访员以及其他人员是否在场也会影响到受访者的行为。在某种情境下，其他在场的人或许有机会听到甚至看到受访者提供的隐私信息。

在调查中，对隐私保护的原则与对保密信息保护的原则是类似的。访员的在场意味着，除了研究者之外，至少访员了解受访者的应答。如果受访者口头作答，访员就直接听到了应答。如果在访问中还有其他人（如其他家庭成员、同事、路过的）在场，在访员与受访者之间的互动中，他们就会听到他们不该听到的，甚至看到他们不该看到的。隐私泄露意味着受访者在接受访问的过程中对隐私的失控。在某些情况下，也意味着一些毫不相干的人知道了他们不该知道的信息。

随着信息搜索变得敏感或者潜在地涉及受访者，隐私问题的影响也变得越来越大。举例来说，正如调查的名称所示，全国药物使用与

健康调查（NSDUH）关注一个极端敏感的议题：非法药物的使用。在某种极端情况下，如果面访是在受访者家里进行，就很难让其他家庭成员不在场。拦截调查以及小组调查（如课堂情景）也很难有什么隐私保护。例如，全国教育进展评估（NAEP）就在样本课堂进行调查，在这种情境下，全班的学生都要参加测试。

隐私的程度，也依受访者而有所不同（参见Beebe，Harrison，McRae，Anderson，and Fulkerson，1998）。在另一种极端的情境下，在门诊或实验室单独空间的自访调查，是考虑到了保护受访者隐私的。对隐私构成威胁的既有家庭成员（例如说到某人的父母中有人吸毒）的在场，也有访员的在场。例如，如果访员是某群体的成员，就很难容许其对该群体有负面态度。

已经有不少新的方法用来提高对受访者的隐私保护。例如，纸版的自访问卷（SAQs）在面访中就已经有较长的历史，其目的就是为了隐匿受访者的隐私。这也是NSDUH采用自访来调查药物和酒精使用的缘起。这种方法在计算机出现以后，就形成了基于计算机辅助自访（CASI）的多种技术形态。在敏感的部门，受访者直接与计算机互动，而不是访员。在语音CASI中，访题甚至都不出现在屏幕上，受访者用耳机听访题，并将应答直接输入计算机。类似的技术，还有从电话调查发展而来的语音互动（IVR）或电话语音CASI，在涉及敏感问题的时候，也增强了对受访者隐私的保护。

5.1.4　沟通的渠道

“沟通渠道”（channels of communication），意指从外界获取信息的多种感知形式。通过视觉、听觉及触觉的不同组合，我们与他人进行沟通；每种组合对我们的理解力、记忆刺激和社会影响的作用不同，从而影响到我们的判断和反应。因此，在如何与受访者沟通访题以及受访者如何将应答传达给调查者的做法上，不同的调查模式有不同的方式。访员引导的调查，主要依靠听觉，即访员大声读出访题，受访者也以同样的方式作出反应。邮寄问卷调查则依靠视觉，即受访者阅读访题，并在纸上写出答案。例如，面访调查可以使用视觉辅助（如出示列有一道访题所有选项的卡片）。CASI可以只是文本形式（视觉），也可以是听觉加文本，或仅用听觉。还有，话语与图片，也是不同的测量工具。因此，对数据搜集方法的另外一种划分，就是看刺激和应答只是使用了话语（如电话模式），还是使用了其他视觉刺激（如图片、视频等）。数据搜集过程引入计算机技术（CAPI和CASI）之后，互联网的使用，也大大扩展了调查中可以给受访者呈现的材料类型（参见Couper，2001）。

早期研究电话沟通的社会心理学研究者，对沟通中由视觉与听觉在不同组合产生的不同效果开展了研究。例如，Short，Williams和Christie（1976）发现，在沟通中，视觉通道对“社会在场”的感知更强烈。“社会在场”（social presence）是一种心理感觉，指主体意识到自己正在直接与另一个完整的行动者接触，并且能够感知到他/她的情绪状态，这代表主体在这一过程中将注意力完全放在交流上（对应多重任务）。因此，社会在场是一种可以在交流中突显另一个行动者主体性的机制。Short，Williams和Christie通过一系列实验还发现，当必须通过观察另一个行动者的情绪状态受访者才能完成组合任务时，听觉通道会加大任务的复杂性。放到调查模式上看，这或许

表明，电访模式中，受访者对访员问题的理解，恐怕不如面访模式全面。在面访模式中，访员的非言语线索也可以被受访者察觉出来。

同样的道理，面访调查中的访员，也更容易察觉受访者不情愿或不清楚的非语言信号，不用受访者明确提出来，访员就可以相应地给以鼓励或澄清。在自访调查中，由于没有这些多重通道的沟通，也就不可能作出这类干预。

早期对调查模式效应的研究，很大一部分关注的是电访与面访两种模式的比较，集中讨论的也是沟通渠道问题。最近，研究的注意力又转移到了视觉与听觉的区别上，一方以邮件及网络调查为代表，而另一方则以访员引导调查及电话自助调查为代表。例如，数据搜集中的首呈效应在视觉模式中更普遍，而近呈效应则经常出现在听觉模式下。首呈效应 （primacy effects），第一个呈现的选项（或者至少接近列表前段的选项）会增加受访选择该选项的几率。近呈效应（recency effects），情况正好相反——会增加对放在列表最后或接近末尾的选项的选择（见第7.3.6节 ）。

对邮件与网络调查模式的研究，也表明视觉形式（访题与应答选项在纸质或电脑屏幕上的版面设计）也会影响受访者给出应答。因此，想要在自访调查中减少测量误差，理解视觉设计与视觉沟通的原则很必要。

5.1.5　技术的应用

最后，问卷调查数据搜集的方法，因技术应用的程度和类型而有所不同（参见Fuchs，Couper，and Hansen，2000）。举例而言，邮寄

问卷调查使用纸版来搜集数据，只要具备读写能力，加上一支铅笔就足够了，不需要其他专门技能和设备。而网络问卷调查，受访者要在网络上使用他们自己的硬件（计算机、调制解调器等）和软件（ISP，浏览器）来运行问卷调查工具。在计算机辅助调查中，访员要使用调查机构提供的技术。应用什么技术以及这些技术如何应用，对调查的标准化（由谁控制设备），访员需要培训的程度和内容（如，CAPI访员在维护与处理笔记本电脑，以及在数据传输等问题上，比统一设备的CATI访员需要更多的培训），调查覆盖率（网络问卷调查需要有条件连通网络；邮寄调查只需要读写能力），以及数据搜集费用等许多方面都会产生影响。当然，运用技术，也可以对复杂问卷进行自动定向、编辑检查等，从而减少某些测量误差。但另一方面，计算机相关设备复杂性的增大，也可能导致其他类型的错误（如编程错误）。

在技术应用中还要留意一个问题，即受访可以操控的程度。一个极端，纸笔SAQ很少会限制受访者。而另一个极端，CASI工具可能会严格限制受访者应答的长度，还会限制应答的顺序，甚至要求受访者只有应答完当前访题才可以继续下一个访题。尽管这样模式下产生的数据可能更加整洁（见第10章 ），但这些约束也可能影响受访者对访题作答的方式。譬如Richman，Kiesler，Weisband和Drasgow（1999）在对多种实验性管理模式的社会期许扭曲的元分析中发现，如果允许受访者回头检查应答，或更改之前的应答，那么，社会期许效应就会减小。

与此相关，人们也越来越多地认识到，使用计算机设施的设计，无论是CAPI、网络，还是其他方式，都会影响访员和受访者的行为，进而影响测量误差。为了改善这些工具的设计，人机交互原理和以用户为中心的设计，越来越多地得到了应用。

5.1.6 上述维度的意涵

在调查研究者看来，如此宽泛且众多的数据搜集工具，其实只代表了几种含义。在说到一种数据搜集方法时，必须对方法本身十分清楚。简单地说某个数据搜集方法（如面访问卷调查，或访员主导的问卷调查）比另一种（如电访或自访问卷调查）更好或更差，而不说明某个方法在上述5个维度上的具体情况，是不够的。譬如网上问卷调查的优势在于减少成本，增加时长，改善测量；不过，却面临覆盖与无应答问题，以及概率抽样框的建构。

与此有关的另一层含义是，很难通过模式比较得出一个普遍的、概括的结论。某种数据搜集方法在某类误差来源上产生的效应取决于各类方法的搭配使用。研究文献还没有涉及所有误差来源的变异性。随着新方法的发展，进行新方法与老方法之间的比较就很有必要。因此，理论对于我们很重要，理论会告诉我们某种方法可能会造成怎样的影响。这些理论既来源于过去模式的相应文献，也来源于对某设计特点或要素的理解。

而且，调查经常综合不同的方法，使用混合模式或杂交设计。正如我们提过的“全国刑事犯罪受害者调查”（NCVS），有一部分面访方式（第一次访谈），余下的部分则通过电话。尽管这种组合可以在整体上缩减费用，同时也会影响无应答、覆盖率与测量误差。同样，像是“全国药物使用与健康调查”（NSDUH），许多部分都采用自访方式（包括纸笔方式，以及计算机方式），并辅之以访员调查。对自访调查，何时做，应该包括哪些访题，诸如此类的问题，在做决定之前，都必须考虑各种方式可能带来的调查误差与费用。在本章后面的

章节，我们会进一步讨论混合模式设计的有关问题（参见第5.4节）。

选择哪一种模式就要具体权衡了。为了满足不同的要求，调查设计常会变得越来越复杂。对某部分最好的方式或可以令某类误差来源最小的方式，可能对另外的部分或另外的误差来源没有帮助。例如，像有声CASI那样直接与计算机进行对话，可能会减少许多测量误差，却也会造成无应答的增多。年龄大的或受教育程度不够的人，可能不愿意通过计算机来回答问题（Couper & Rowe，1996）。还没有哪一种数据搜集方法对于任何情况都是最好的。选择某种方法，必须客观考虑具体的调查情境与已有资源。

5.2　选择合适的方法

如果没有适用于所有条件的理想模式，那么，对于某项具体研究来说，又如何选择恰当的方法呢？做决定之前，有几个问题必须要权衡，必须考虑不同种类的调查误差，还有费用（广义地看，包括了后勤、人员、时间和其他组织问题）。有时候，选择显而易见。例如，通过邮寄方式调查文化程度就很不明智，因为文化程度低的人可能根本无法理解访题。同样，文化程度也很难通过电话来调查。那么使用什么方法可以测量文化程度，既不用受访者阅读材料，也不要测评他们这方面的能力呢？出于这种考虑，面对面交流可能比较恰当。全国教育进展评估（NAEP）曾通过测试学生的阅读与数学知识来衡量学校的绩效。在学校做这样的测试，就很合适，因为在学校，学习是大家最关心的事情。

有时，很容易排除一些备选方法。例如，如果一个人想估测互联网的普及程度，在网络上作调查就毫无道理可言（尽管许多研究者一直都在这么做）。同样，要估测多少人在医疗问题上遭遇过麻烦（如BRFSS做的事），仅仅使用医疗记录也很容易遗漏许多信息。

对于大部分调查设计来说，调查研究者必须在多种方法之间反复权衡，考虑不同因素的相对重要性，才能作出最终决定。如果说，相对于测量误差与费用来说，覆盖性误差的重要性起了决定作用，那么，在方法的选择上可能就会完全不同。

尽管可选的范围很广，方法之间却有着显而易见的逻辑关系。例如，电话调查经常被认为可以替代面访，大概就因为这两种方法都会用到访员，并且抽样框覆盖情况大致相似。因此，行为风险因素检测系统（BRFSS）虽然使用的是电话调查，实际上也可以用面访来取代（还会提高家庭总体的覆盖率），只是费用更多一些。NCVS在首轮调查中就使用了面访，之后的几轮，大部分都改为电话调查。

邮寄调查可能是电话调查的一个可替代方式，条件是，需要样本列表中既有电话号码，也有住址信息。网络调查可以替代邮寄调查、街上拦截方法，甚至电话调查技术，但也有它无法替代的，比如面访调查。CES之所以经久不衰，拥有商业的目标群体，就因为除了最初基于客户姓名和地址的信息之外，还获得了其他的信息（比如，电话号码、电子邮件地址），可以为多种模式提供单位样本。

如果在一个多维空间，用我们前面提到的几种维度列出所有的方法，我们会看到有些方法之间联系更紧，有些则差异较大。一般说来，方法之间相似性越大，彼此可替换的可能性就越大，这些比较都来自于关于模式比较的文献。

5.3　不同数据搜集方法对调查误差的影响

我们对不同数据搜集模式之间的相对优势与不足的认识，大部分来于模式的比较研究。一般说来，这些研究都采用了实地实验方法，将单位样本的一部分随机安排到一种模式，另一部分则安排到另外一种模式。还有几个元分析研究，也能加深我们对模式效应的理解（例如，Goyder，1985；de Leeuw & van der Zouwen，1988）。做这类模式比较研究很有难度，而且由不同设计提供的信息经常也有局限。但是如果要对各种方法的相对价值作出评判，理解这些问题就相当重要。

这里，我们无法对所有可能的模式比较进行全面和详尽的论述，只能集中讨论关键的几组比较，包括面访与电话，电话与邮寄，以及邮寄与网络。

5.3.1　测量模式的边际效应

因模式不同会造成调查估计也不相同，这有许多原因。例如，使用的样本框不一样（如，针对电话样本的RDD框和针对面访的区域概率抽样框）。抽样框总体也会不一样（如，在邮寄与网络模式比较中，不是所有样本都上网）。无应答率也会随模式不同而不同。因此，要明确模式之间差异的特定来源经常非常困难。对于一些比较，可以做到这一点——比如，在面访与电访进行比较时，可以先排除没有电话的家庭，这样因家庭总体覆盖不同而产生的影响就可以排除了。同

样，为了排除抽样框总体不同而产生的混淆，邮寄与网络调查的比较可以限制两种模式的样本为都上网的人。表5.1，出自Groves的研究（1989，图11.1），表明了在对电访与面访两种模式比较中，应当出现的设计问题。相似的问题同样也适用于其他模式之间的比较。

表5.1　比较面访与电访问卷调查的设计问题

设计特征	重要问题
样本框	同样本框,还是电话用户样本框?
访员	同一批访员？相同的招募与培训？相同的经验?
督导	电访是否集中进行？联系同样的督导?
应答规则	应答选择程序一致?
问卷	统一的问卷？也有可视化辅助?
回访规则	一样的规则？同样的强度?
对待拒访的方式	同样的努力程度?
计算机辅助	使用 CATI/CAPI?

模式比较的一个策略是，把每种模式作为一个特征组，然后比较不同特征组之间的净差异。这个策略强调的是实践性问题，即不同调查模式是否会得到不同的调查结论。对最佳模式的选择（如面访与电访），通常不需要对每个方面都做严格的比较。要做的是，针对每种模式，进行访谈方式、样本框、设计工具等的优选。当认为调查可以用不同的模式彼此替换时，经常会使用这个方法。比较的焦点是不同模式得到的结果之间是否相似还是不同，以及差别背后的特殊原因是什么。“当前就业统计调查”（CPS）在从纸笔转为计算机形式的时候，就用这个方法评价了模式改变带来的整体影响，以及修改后的CPS问卷是否完备（Cohany，Polivka，& Rothgeb，1994）。

模式比较的另一个策略，是专注于理解两种模式背后不同的机制，通常使用同样的抽样框。它试图隔离某种特殊的因素（例如，沟通渠道），在最初的筛选之后（以此排除覆盖率和无应答的差异），

随机将受访安排到不同的模式下。另一个例子是评价技术对调查的影响。将采用纸笔问卷的电访转变为CATI，但对工具不进行相应的改进，也不应用高级CATI功能。这个方法经常要在设计上作出妥协，例如，不能根据各种模式独有的优势来再次设计问卷。有时，这些研究背景是长期调查。在长期调查中，通常第一轮调查会采用面访，后面的调查就改为其他的模式。模式比较因而也会因不是同一轮调查之间的差异而造成差异的混淆。

Hochstim（1967）对面访与电访以及邮寄问卷三种模式的研究

Hochstim（1967）的报告是最早也最有影响力的一篇关于模式比较的研究。

研究设计：面访访员访问了加州Alameda县的350个小区，共计2 148户，97%的住户接受了调查。之后，将受访者随机安排到面访、电访以及邮寄问卷三种模式之一。使用同一种设计，进行两个不同的调查。如果最初安排的模式无人应答，就换另外一个模式。第一个调查涉及一般的医疗、家庭和人口统计学等信息；家庭所有成员都在其中。第二个调查涉及流产手术，受访者为20岁以上的女性。之后，访问医疗相关机构，以获得有关的记录信息。

研究发现：面访模式有最高的首次应答率，首次未完成的样本会安排到其他模式下。面访完成样本的费用也最高，电访和邮寄问卷调查的费用只是面访的12%。三种模式之间不存在人口变量差异，也很少出现实质性的差异。在邮寄问卷模式中，遗漏数据的项目更多，对社会不赞许行为（例如，饮酒）的回应也更多。比较调查获得的流产数量与医疗记录数据，结果没有差异。

研究局限：实验组不是单一模式，而是一个主要模式与次要模式的组合，模式差异容易混淆。使用一个县作为目标总体，只关注健康和医疗状况，其结果只能有限地推广到其他类型的调查。

研究意义：跨模式之间取得了相似的结果，每种模式都各有千秋。对不同模式费用和社会赞许敏感度的问题，在后来的研究中都得到了检验。

文献中关于模式比较的研究，有从实际出发的（比较各类特征），也有从理论出发的（试图分离某设计特点产生的效应）。如何混合和匹配各类数据搜集方法，发挥其独特的优势以达到最佳的效果，还没有一定之规。然而，对一般的调查特征（例如，访员的行为，问题格式，覆盖总体），都已有研究。综合几个这类研究的结果，我们可以根据调查误差来源与费用总结出各类模式的相对优势与不足。

接下来的几个章节，我们将通过实验证据，说明模式之间在调查误差来源与费用等方面存在的差异与相似之处。

5.3.2　模式选择对抽样框和抽样设计的影响

有什么样的抽样框可用，经常会影响到调查方法的选择。例如，如果抽样框有每个样本的邮寄地址，那么选择邮寄调查就说得过去。如果抽样框有每个样本的电邮地址，则可以采用网络调查。反过来，

也可以用设想的数据搜集模式来选择合适的抽样框。如果希望对一般人群做电访，就一定会用到电话号码框。数据搜集模式和基本抽样策略的选择，通常是同步完成的。

几乎所有的调查，只要使用区域概率抽样框都会从面访开始，尽管后面几轮可能会改为较便宜的数据搜集方法（如果需要多次与受访者访谈）。电话号码抽样框几乎专门被用于电话访谈或IVR。

de Leeuw和van der Zouwen对电访和面访调查数据质量的元分析研究（1988）

1988年，de Leeuw和van der Zouwen报告了对31个面访与电访比较的元分析研究（例如，对已发表的研究在统计上再做综合和汇总）。

研究设计：重新整理从1952年到1986年总计28个对电访和面访调查比较研究的数据，包括已发表在杂志上和未发表的论文。质量测量的内容包括对记录数据、有无社会赞许偏差、访题缺失数据、应答的多少（开放性访题）、模式之间的相似性，和样本无应答率等方面的比较。

研究发现：面访调查的平均应答率为75%；电访调查为69%。模式之间在测得的社会赞许偏差与项目缺失数据的比率区别很小，还发现，在当时，受访者一般更倾向于接受面访。这些证据表明，随着时间的发展，模式本身之间的差异性会降低，所以后来研究发现的模式效应更小一些。

研究局限：如所有元分析研究一样，该发现也只限于当时所能得到的一系列研究。没有包括计算机辅助效应的研究。绝大多数研究没有对无应答和测量误差作区分。一些面访调查中包括了无电话样本，这就混淆了覆盖性差异与测量差异。因此，这里所谓的“数据质量”还包括了观察和无观察误差。所有研究都是1986年以前做的，那个时期，电话调查还不那么普遍。

研究意义：研究在广泛的实际领域，对电访调查的普及提供了支持。

方法选择对样本设计常有间接影响。因为面访费用昂贵，这种数据搜集模式几乎总会用到整群抽样，即使会降低样本有效性。相反，如果邮件或网络调查也使用整群抽样设计，就不会节省太多费用（见第3章 和第4章 对不同抽样设计与抽样框的讨论）。

关于数据搜集模式对抽样设计影响的最后一个考虑是：如何从样本家庭选取受访者。这个选择最好由访员来做，不要让受访者自己选择。在面访调查中，我们需要所有家庭成员的列表，然后用随机方式抽取样本，如此，其覆盖性误差往往比电访要小。目前，在电话访谈中，也可以有几种不同的程序，使得误差降到最低（见第3章 和第4章）。但在自访问卷调查中，要想尝试同样的策略就很难做到。同样，需要筛选受访者的调查（例如，有年龄限制），也最好由受过培训的访员来做。如此，对绝大多数自访问卷调查而言，研究者如何控制究竟是哪些人在应答，就成为亟待弥补的不足。对这个议题，还有待大量的研究。

5.3.3 模式选择对覆盖性的影响

数据搜集方式和抽样框的选择，经常相互关联，因而，更多的问题在于方式选择与目标总体覆盖性的关系。例如，RDD样本框就是电话号码列表，对调查家户人口而言，通过电话进行首次联络是最好的方式。电话簿抽样框的覆盖性一般都比较糟糕，好在通过它通常可以获得邮寄地址。在RDD生成号码之后，反向查找姓名和地址目录，通常能获得一半的邮寄地址信息。电子邮件地址抽样框只能在网络调查中派上用场，除非还有其他信息。不过，只要可以获得抽样框信息，这些方式就会越来越有吸引力和可行性。然而，不应忽视的是不同抽样框的覆盖性特征。

从家户人口的覆盖性出发，区域抽样框与面访的结合可以被认为是其他方法无法媲美的黄金搭档。即便如此，考虑到费用、效率或覆盖性等因素，通常还是会以某种方式限制这类面访的总体。例如，经常将抽样框限定为普通居民（军人除外）、非机构性的组织（监狱、医院和其他机构除外）、家户（无家可归者、旅客除外）、美国大陆人口（阿拉斯加、夏威夷和其他领域除外）。总而言之，总体的某些亚群体比其他人更难进入抽样调查中，往往因为费用或效率的原因，这些人群就被排除在外了。我们的绝大部分调查样本都限于普通家户人口。

美国的电话覆盖率（拥有固定电话家户的比例）在过去的几十年里已经超过了90%（1980年92.9%，1990年94.8%，2000年94.5%）。不过随着仅有手机的家户数量的快速增长，到2008年的早期，固定电话的比例已经下降到了80%，有17.5%的家户仅有无线接入，即电话覆盖率为97.5%（Blumberg and Luke，2008）。电话没有覆盖到的人群与

被多种方式覆盖的人群，在多个维度上都有差别，特别是其社会人口变量（Blumberg，Luke，Cynamon，and Frankel，2008）。对某些调查而言，例如针对刑事犯罪受害者、吸毒、失业，以及领取救济人群的调查而言，就必须考虑这个问题。

对于邮寄调查，美国没有通用的人口列表。或许有一些替代，如投票人登记表就经常被用来针对某主题做州域范围的邮寄调查。显然，不是所有居住在某州的成年人都会登记为投票人；2006年11月大约只有72%的合法美国成年公民登记投票（Bureau of the Census，2006）。因此，邮寄调查更多用来调查已有样本框或列表的特定人群。覆盖性（即非覆盖性误差）在很大程度上取决于使用的特定列表。（在保留人口登记的国家，这方面很少会成为问题，因为全体居民的邮件地址理论上都可以获得。但即便有这样的列表，也并不意味着它就是完整准确的。）

至于网络或互联网调查，最近几年，美国成年人接入互联网的比例迅速增长，从1995年的低于15%到2000年的50%，再到2008年的75%（见图5.3）。与其他接触方式（如，邮件、电话）不同，接入或使用互联网技术，并不意味着可以用技术实际接触到样本个体；大部分网络调查在开始时仍然通过电子邮件与样本个体打交道，也没有构建出可以对互联网使用者总体进行抽样的抽样框（参见Couper，2000，第3.4.3节）。除了接近25%的人群无法覆盖以外，如果要将结果推广到所有家户人口，还必须考虑接入网络人群与没有接入网络人群之间的区别。美国电信管理局（NTLA）在他们的系列报告中提到的“数字鸿沟”（digital divide），就是这两类人群差别的证据。举一个例子，根据2008年5月发布的皮尤互联网项目报告（Pew Internet Project Report），在18—25岁的人群中，有90%的人使用互联网；而

在65岁及以上的人群中，只有35%的人使用互联网。同样，在有大专及以上学历的人群中，有91%的人使用互联网；而在高中及以下学历的人群中使用互联网的比例只有44%。使用与不使用互联网的人群之间的差别，就是所谓的“数字鸿沟”，且不只限于这些人口特征（参见Robinson，Neustadtl，& Kestnbaum，2002；Couper，Kapteyn，Schonlau，and Winter，2007）。

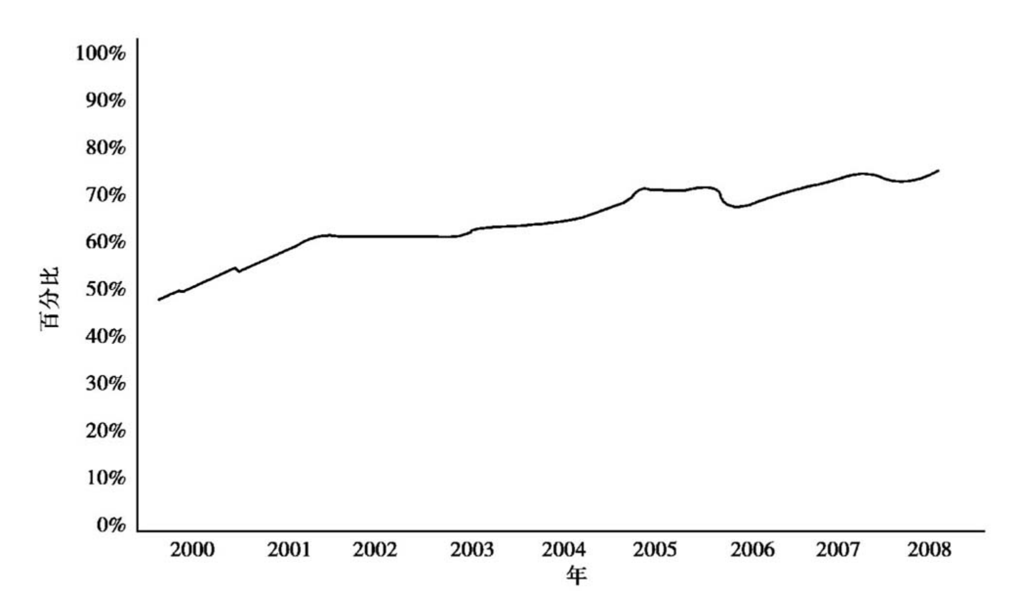


图5.3　美国成人使用网络的百分比，2000年第一季度至2008年第三季度

（数据来源：Pew Internet and American Life Project Surveys，2008.）

对既有的调查，如CES，美国政府机构使用了他们掌握的样本框，有准确的姓名和住址。不过，电话号码一般不可信。因此，许多已有调查，开始时都用邮件进行联系。选择电话调查方法的，一般也必须将电话号码与地址抽样框进行匹配。联邦政府以外的调查研究者，会依靠商业调查来获得抽样框，有些可以提供电话和邮寄地址，因此可以采用电话和邮寄模式。

所以，要选择或评价一个数据搜集模式，就需要考虑不同模式的覆盖属性。事实上，通过电话无法找到的人，很可能是穷人和失业者。如果要做一个关于失业或领取救济方面的调查，使用电话调查模式显然不合适。然而，对于政治问题的调查，就不用过于考虑覆盖率。对于后者来说，在选择模式时，针对覆盖性要着重考虑这样几个因素：

1）速度有时是根本（通过政治民意调查产生估计值的半衰期尤其短暂）。

2）对调查的使用者来说，高度精确可能并不是最重要的。

3）对可能参与选举的人来说，推论经常因电话所有人的不同而变化。

由于各种数据搜集方法的一般目标总体，其覆盖性总会变化，因此从事调查方法的研究者需要对此展开评估。新方法的新抽样框（比如，网络调查）随着时间的推移，会逐步建构出来。为了给调查研究者提供实际的指导，有必要对其覆盖性进行研究。

5.3.4　模式选择对无应答的影响

数据搜集方式选择也会影响无应答比率和无应答偏差，有大量关于模式差异的研究都关注于此。除了对两模式、三模式的比较之外，还有几个针对应答率而展开的元分析，其中有同一模式下的（如Heberlein，Baumgartner，1978；Yu，Cooper，1983；Edwards，

Roberts，Clarke，DiGuiseppi，Pratap，Wentz，& Kwan，2002；de Leeuw & de Heer，2002），也有跨模式比较的（Goyder，1985；Hox & de Leeuw，1994）。例如，Goyder（1985）整理了112份面访调查和386份邮寄调查的应答率和其他信息，并对此作了元分析。他发现，如果控制了接触次数、样本类型和资金赞助行为等变量之后，“对邮寄调查和访谈的应答率，基本上都会降到一个差不多的水准，但访谈法成本要高得多。”（Goyder，1985，p. 246）。说到差别，平均来说，面访调查的应答率比邮寄调查要高，但这些差别通常反映在尝试接触的次数和其他等变量上。Hox和de Leeuw（1994）对45个研究进行了元分析，仔细比较了邮寄、电话和面访获得的应答率。他们的结论是，从平均水平上看，面访调查应答率最高，接着是电话调查，其次是邮寄调查。同样，Groves和Lyberg（1988）也提到，电话调查的应答率往往不如面访调查。

是什么导致了这些差异呢？第6.7节提到了访员在传达调查需要、澄清问题及对样本表达关注等方面所发挥的优势。这种优势的有效性与调查方及访员的经验密切相关。面访调查的高应答率，很多是由访员在作自我介绍时让对方感受到了高度的可靠感。电话调查，只依靠听觉，就无法给人呈现实实在在的感受（如，身份证、官方签署的信件）。缺少这些，访员感召受访者的能力受限了。在自访问卷中，没有这样的人际感召，受访者对调查的重要性和合法性了解就更加缺乏。在面访中，访员可以适当地劝服，用以排除受访者的顾虑，但自访问卷调查就没有这样的机会。

对访员指导下的调查，还没有证据表明技术的使用（比如，CATI或CAPI）会影响应答率。Nicholls，Baker和Martin（1997）回顾了计算机辅助方法与纸笔方法的比较研究，发现这两种方式在无法联系与

拒访率上没有显著差别。同样，CASI技术也对无应答率作用很小；既然它们通常都作为访员指导下调查的一个部分，这个结果也在情理之中。

然而，对自访调查而言，纸笔方式往往比使用电子设备（电子邮件和互联网）获得更高的应答率。尽管有一些例外，最近的两个元分析（Lozar Manfreda，Bosnjak，Haas，and Vehovar，2008；Shih and Fall，2008）清楚地显示了邮寄问卷调查相对于互联网调查的优势。两者的差别到底是来自于方法的不同，还是因为目前还没有更好的、已经经过检验的策略来提高网络调查的应答率，还有待进一步探究。

一般来说，还不清楚到底是因为模式之间存在直接影响应答率的固有差别，还是因为不同的模式在实施中有许多方面存在不同，从而造成了这种差异，比如与受访者接触的次数、激发合作的方式、鼓励和其他劝服的表达技巧，以及合法性材料等。例如，Cook，Heath和Thompon（2002）对68个网络（或电子邮件）调查进行了元分析，发现联络次数、单独个别联络和提前联络都与更高的应答率相关（Shih and Fan，2008；Couper，2008b）。同样，这些变量也可以影响普通邮件以及电话和面访调查的应答率。

不同的数据搜集方法，得到的无应答信息量也不同，比如能不能确定单个样本是不是真的无应答，如果是无应答，是什么原因造成的（见Dillman，Eltinge，Groves，& Little，2002）。例如，在邮寄调查中，很难在无应答中区分不合格单位样本（比如地址错误）。有一些问卷被退回的理由很明确，即无法投递（PMR或者通过邮递员），有一些有完整的回复；即使如此，也很难讲问卷是不是真的就投递到了目标样本。同样，电话调查中的多次“无人接听”，可能是这个号

码不对，也可能是调查期间主人不在家。在面访调查中，访员通常可以通过观察来确定被选家户是否符合调查要求。

对于无应答的产生过程，是否可以获得丰富的信息，也因调查方法的不同而不同。在电话调查中，几乎没有信息能够说明无法联系到的可能原因是什么，或者能用什么方法来解决，只能反复地拨打尝试。通常，电话调查的初次交流要比面访调查简短得多（Couper & Groves，2002），因而也无法得知更多的原因及相关信息。邮寄调查对无应答过程一般也没有多少信息。不过，如果预先设定好受访者，网络调查有时候倒可以在一些方面提供详细的情况（Bosnjak & Tuten，2001；Vehovar，Batagelj，Lozar Manfreda，& Zaletel，2002）。例如，研究者能够区分出完全上不了网的人和登录到调查页面却没有完成问卷的人，以及完成了大部分问卷访题却不再继续的人。

一种方法在某个方面对无应答有优势，并不说明在其他方面也一定有优势。例如，邮寄调查虽然经常出现邮寄错误，但与访员指导的方式相比，却是与人、与家户或机构联系的一种最经济的途径。因为缺少直接接触，获得样本合作的能力就有限。电话访谈的费用也要比面访低很多，各类访问工具（如，留言机，来电显示）的使用，也降低了电话调查的接触率。面访调查需要的费用会更高，但接触后却比其他方法更容易获得合作。由于访员在场，他/她可以根据样本的特殊顾虑，运用不同的说服策略（Groves & Couper，1998）。Groves和Lyberg（1988）提到电话调查的无应答比例很大部分与拒访率相关。因此，说服不情愿的人参加调查，取决于媒介的丰富性（如书信的鼓动作用就很有限），也取决于受访者的反馈（根据不同的反馈，才能作出不同的应对方式）。

总之，数据搜集方式不同不仅会有不同的无应答率，且无应答的原因也各不相同。对无应答偏差来说，后者的意义可能更为重要。例如，邮寄调查比访员指导的调查造成的无应答偏差更大，因为样本个体在他/她确定是否要参加调查之前，可能已经看过调查的内容。要不要参与，可能取决于他/她对调查内容的反应，或也取决于他/她对调查感兴趣的变量所持的价值观。在访员指导的调查中，调查的内容经常可以用模糊、委婉的说法来掩饰（如把毒品或性方面的调查说成“健康与社交生活”），以免调查对象察觉出调查的主题。显然，目前对调查方法的研究，还有需要挖掘各种方法之间无应答差异的机制。

5.3.5　模式选择的测量质量

模式也会影响搜集数据的质量。我们这里集中看数据质量的三个方面：数据的完整性、应答因社会期许偏差所产生的扭曲程度，以及应答受其他应答效应影响的程度。“社会期许偏差”（social desirability bias）指用偏爱的方式来表达自己的倾向。如果受访者夸大了社会的期待（如投票）并隐瞒社会的反对（如使用毒品），就构成了社会期许偏差。数据搜集的方式会影响社会期许偏差的水平。“应答效应”（response effects）是指调查因访题用词的精确性或应答选项的排列顺序，甚至访题的顺序等特点造成的测量问题。例如，如果问题换一种说法，或者两个相关的访题顺序颠倒一下，答案就有可能发生变化。

Toruangeau和Smith对回答敏感访题的模式效应研究（1996）

1996年，Toruangeau和Smith比较了用CAPI，CASI和ACASI三种方法获得的性行为数据。

研究设计：随机实验，在样本调查地区，对643名年龄从18岁到45岁，来自芝加哥32个区域的受访者进行调查，应答率为56.8%。每个区域内部，访员只能使用CAPI，CASI或ACASI其中的一种调查方式。关键的报告数据是非法使用毒品和性行为的水平，以及样本无应答率。

研究发现：ACASI和CASI方法获得的药物滥用和性行为水平一般较高。例如，受访者在ACASI与CASI中报告服用大麻的比例，分别比CAPI高出了48%和29%。报告有肛交经验的在ACASI中比CAPI中要高出421%，CASI则高出204%。用ACASI和CASI，与CAPI相比，男性报告出更少的性伴侣，而女性则报告出更多的性伴侣（这表明在自访答的模式下，社会期许效应更小）。无应答率在三种模式中没有明显差异。

研究局限：没有办法评估不同模式获得的结果是否有效。不过，这里用到的大部分访题都有社会期许行为，会有低报现象。样本主要是年轻、受过教育的城市居民，这些人比其他人群在使用CASI和ACASI工具时，会更容易上手。

研究意义：这项研究用实验方法支持了自访模式能给受访者提供更可靠的隐私性，并且对社会反对行为的报告也会更加准确。因此，当测量人们对待敏感问题的态度时，许多研究都使用自访答的方式。

考虑到数据的完整性，自访答问卷（如邮寄问卷）中无回答的访题比访员指导下的调查更多（Tourangeau，Rips，& Rasinski，2000）。这可能有三个原因：（1）受访者不理解访题（访员不在场，无法提供帮助）；（2）受访者没有按照问卷的要求来做；（3）受访者不愿意作出回答（访员不在场，无法说服他们）。尽管也有例外（如de Leeuw，1992），但许多模式比较的文献都列出了大概相似的趋势（如，Brøgger et al.，2002；Van Campen，Sixma，Kerssens，& Peters，1998；O'Toole，Battistutta，Long，& Crouch，1986）。在互联网调查中，对缺损值的了解还不是很清楚，一方面可能是由于互联网调查可以设计为许多种形式，例如接受缺损值（正如在邮寄问卷中一样），鼓励受访者尽量填答（就像访员指导下的调查一样），或者在继续下一步之前再请受访者应答。因此，互联网调查中的缺损值更多的可能是设计的后果，而不是调查方式的影响。

电访与面访在缺失数据比例上的差异，很少存在争议。Groves和Kahn（1979）发现，在整体上，电访的缺失数据比例更高，这一结果与Jordan，Marcus和Reeder（1980）以及Körmendi和Noordhoek（1989）针对收入访题所获得的结果一致。Béland和St-Pierre（2008）报告的结果中，在多个涉及健康的访题上，CATI比CAPI的缺损值比例更高。还有一个研究报告得到了差不多的比例（Aquilino，1992；Aneshensel，Frerichs，Clark，& Yokopenic，1982；Dillman，1978；Hochstim，1967）。电访的自动值机（即使用电子语音，而非由访员询问，有利于受访者提供敏感信息），即便无法像面访那样可以使受访者确信调查的合法性和保密性，也会对数据缺失有所弥补。之前曾经提到过，技术的使用会影响选项数据缺失。计算机辅助方法往往容易降低缺失数据的比例，与纸质调查相比，基本可以排除漏题现象，因为这类工具会精确地执行预先编好的程序。

正如Martin，O'Muircheartaigh和Curtice（1993）所发现的，与纸质调查相比，CAPI可以明显地减低选项缺失数据的比例，但是对“不知道”和“拒绝回答”的应答比例，却不存在差异。

对应答完整性的测量还可以看开放题的应答长度。“开放题 ”（open-ended questions）允许受访者用自己的话来表述对问题的看法。例如，许多调查会询问受访者的职业，然后由受过训练的编码员对答案进行分类。相反，封闭型访题只要求受访者从题目已列出的几个选项中选择自己的答案。在面访与电访的比较研究中，Groves和Kahn（1979）的研究发现，对于开放题，面访得到的可编码应答明显更多。他们认为这是因为电访速度比较快，缺乏可利用的非言语线索（也见Groves，1979）。Körmendi和Noordhoek（1989）得到了相似的结果。关于自访与访员指导下的调查在这一维度上的比较还不多。一方面，因为邮寄调查中受访者没有太大的时间压力，他们想写多少就可以写多少，因此或许他们的回答会更长。另一方面，访员通过盘问，或许可以从受访者那里成功获得更多的信息。De Leeuw（1992）对4个开放题的回答进行了比较，发现对其中的两道访题来说，邮件与访员指导方法之间不存在差异；另外两题目，有一题邮寄调查回答得更长，另一题访员指导下回答得更长。Bishop，Hippler，Schwarz和Strack（1988）发现，在“对一份工作最看重什么”的问题上，与电访相比，受访者在自访答表格里更愿意给出多个答案。他们的解释是，在自访条件下，受访者无法对每个回答做出必要的澄清与判断。对纸质调查与计算机化调查的比较发现，计算机辅助访谈对开放题的回答没有显著影响。关于行业和职业题的回答，CAPI（Bernard，1989）和CATI（Catlin & Ingram，1988）与同样的纸笔方式，在比较中没有发现长度或质量（可编码性）上的差异。Denscombe（2008）与

Deutskens，De Ruyter和Wetzels（2006）的研究指出，互联网和电邮调查，在题意明确的开放题上的应答与其他方法没有什么差别。

总之，比起自访答调查与纸质数据搜集方式，访员指导下的和计算机辅助调查看上去可以减低缺失数据的比例。与电访相比，面访对于开放题好像也能带来更丰满的回答。

文献提到的最为明显的模式效应中还有社会期许效应，即指受访者喜欢将自己的应答指向社会期许的方向。例如，没有参加投票选举的人可能会称自己投过票，吸烟者可能否认自己吸烟，有族群偏见的人或许不会表露对待少数族群的真实态度。受访者对那些会令自己尴尬的访题经常回避，不愿意回答，数据搜集的方式看上去会影响他们是否愿意承认那些不受欢迎的态度或行为。一般说来，访员在场会增加社会期许效应，即夸大社会期许的行为，比如参加投票选举，参加教堂活动，锻炼身体，健康饮食等，并隐瞒社会反对的行为，如吸毒、酗酒、乱性之类。因此自访答方式通常比访员指导方式有利于减少社会期许效应。

图5.4列出了两项大型研究针对非法药物滥用的数据，分别来自面访和自访答问卷数据之间的比较。这两项研究都问到了受访者在几个时间段内（上个月、去年、一生之中）对几种非法药物的使用情况，包括大麻和可卡因。图中数据是在自访答与访员指导下，受访者报告服用非法药品的比例。例如，Turner，Lessler和Devore（1992）发现在自访答情况下，报告上个月曾用过可卡因的比例是访员指导下的2.46倍。Schober，Caces，Pergamit和Branden（1992）也得到了相似的结果。

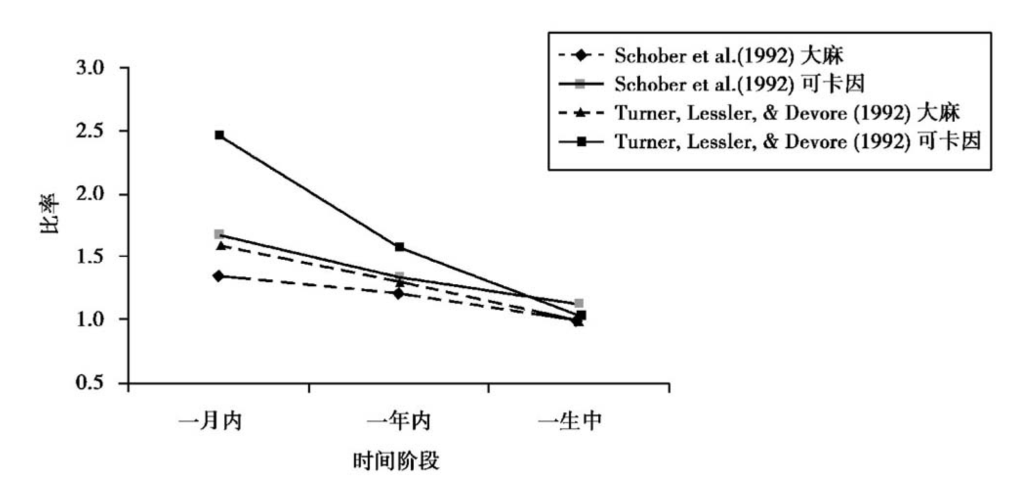


图5.4　在自访答与访员引导的问卷中，受访者报告的在不同时间段内使用毒品的比例

在访员指导的调查中，电访在应付敏感信息时看上去比面访效果更差，从数据来看，表现出了更强的社会期许偏向（如，Groves & Kahn ，1979；Henson ，Roth ，& Cannell ，1977；Johnson ，Hoagland，& Clayton，1989；de Leeuw & van der Zouwen，1988；Aquilino，1992）。但也有几项研究得出了相反的结果（如，Sykes & Collins，1988；Hochstim，1967），还有人发现面访与电访之间没有差异（如，Mangione，Hingson，& Barrett，1982）。这里需要特别指出de Leeuw与van der Zouwen的那项研究，它是一项元分析，对大量的模式进行了比较研究。

许多研究已经发现了不同模式的应答效应差异。应答效应指调查中的许多测量方式的效应，如访题顺序的效应（应答会因为提问的顺序而发生变化），还有应答顺序效应。“应答顺序效应”指如果应答选项呈现的顺序一旦发生变化，应答也会因此改变。在许多例子中，受访者往往倾向于选择第一个或第二个呈现出来的选项（首呈效应）。在另一些情况下，受访者似乎更愿意选最后一个选项（近呈效应）。Bishop，Hippler，Schwarz和Strack（1988）得出一个结论，

与电话调查相比，在自访调查中，访题顺序效应和应答顺序效应明显要少，然而，每种数据搜集模式都会产生访题形式效应和措辞效应。访题顺序效应和应答顺序效应之所以在自访调查中表现得不突出，是因为受访者可以在回答问题之前纵观整套访题，因此会减小这种效应。

不过，应答顺序效应的方向似乎部分取决于数据搜集的方法。如果问题是说出来的，那么最后一项答案更容易被选出来；如果问题是书面呈现出来的，则第一个选项容易被作为应答（Schwarz，Hippler，Deutsch，& Strack，1985；Tourangeau，Rips，& Rasinski，2000）。关于问题的顺序，尽管邮寄调查不能严格按访员指导的调查那样依序进行（因此会减小访题顺序效应），但访题放在同一页还是不同页上，可以一下子看到还是不能（Schwarz，Hippler，Deutsch，& Strack，1991），可能会增加或减少这类相关的情境效应（context effect）。同样，对网络调查而言，一份可以上下滚动的调查，能让受访者看到所有访题，比依序呈现访题的方式更容易减少这类顺序效应。另外，从这些例子中，我们可以看出区分不同模式的重要性，特别是在特定模式下设计和使用测量工具时。

对不同模式的其他应答类型，调查过的还包括默许效应（acquiescence）（不管问题的内容是什么，总是一个应答的倾向）和极端效应 （extremness）（选择位于标尺两端的应答倾向）。这些效应经常很难从社会期许效应和应答顺序效应中抽离出来。因此，这些结果一般都混杂在一起，一些研究发现电访比起面访（如，Jordan，Marcus，& Reeder，1980；Groves，1979）以及邮寄调查（如，Tarnai & Dillman，1992），有更多的默许效应和极端效应。还有人（如，de Leeuw，1992）发现模式之间不存在这方面的差异。

到这里，我们已经简要地回顾了一些测量效应，并对几种数据搜集模式在这些维度上进行了比较。但如何比较那些最新或刚出现的数据搜集模式呢？一般说来，听觉模式（如访员指导的，IVR，语音-CASI）是按照顺序把访题传递给受访者。受访者必须按照预先安排好的顺序，回答完一道访题之后，才能回答下一道访题。在以视觉支配的模式中（如邮件，纸版SAQ），受访者不必再受访题顺序的约束，所有访题都在一份问卷中，随便用什么顺序来回答。因此，我们认为在视觉模式中，情景效应（如访题顺序和应答顺序）会更少，有一些研究也证明了这一点。然而，互联网调查既可以设计成类似纸版问卷的样子（如只有单一滚动的HTML格式），也可以设计成访员指导调查的样子（依序一次只呈现一道或者几个访题）。滚动的设计产生的情景效应更小。因此，情景效应不必是一种模式本身的特点，更多依赖于模式工具的设计。

到目前为止，我们的焦点都放在模式之间的测量误差差异上。进行模式比较，就需要搜集各类模式的可比数据。不过，我们也可以换一个思路，看不同模式提供的测量机会差异。例如，选项卡（show cards）（每张卡片列有不同的应答选项），或者其他的视觉辅助工具（如图片）在面访调查中就很好用，但如果用于电话调查就麻烦了。选项卡会影响应答效应，可以帮助受访者进行回想与确认，因此，在许多领域（如杂志的可读性、广告测试、药品说明，等等）得到应用。对传统调查测量方法进行的其他补充，因模式不同也有差异。譬如面访可以包括观察法（如对住房、生活小区等进行观察），还可以有身体测量（如身高、体重等）。面访数据搜集方法，还可通过受访者对其身体样本（头发、尿液、血液）进行数据搜集，也可以对他们的环境（放射性物质、水质等）进行数据搜集，但在其他数据搜集模式下，这些几乎无法做到（如，Etter，Perneger，& Ronchi，1998；

Boyle et al.，2007）。此外，计算机辅助数据搜集方法，可以对应答倾向和其他数据进行测量（如计算呈现题目和受访者给出应答之间的时间），用纸质方式则很难做到。同样，由于计算机辅助访谈（CAI）有能力将题目和选项进行随机化处理（既不影响测量，还可以降低访题和应答顺序效应），因而也促进了CAI的应用。

5.3.6　对成本的影响

调查费用通常有两类：固定费用和可变费用。固定费用（fixed costs）指不管样本大小多少，调查所需的不变成本。例如，问卷的开发、测试与编程都属于固定费用；这些工作的成本与样本大小无关。可变费用或计件费用（variable or per-case costs），指用在与受访者联络、访谈、跟进无应答受访者等方面的成本，随着样本数量的变化而变化。如果只讨论数据搜集的不同方法，而不讨论相应的成本，就只能算是玩忽职守。不同数据搜集方法的成本，取决于很多实际情况的细节。例如，本地小型调查比全国大型调查而言，访员的交通费用就小得多，这样的方式也适用于面访与电访之间成本的比较。同样，反复拨打尝试的数量和跟进邮件的次数，也会影响邮寄与电话调查的相对成本。邮寄与互联网调查费用的比较也依赖于样本的大小，互联网调查的固定费用占了相当大的比例，邮寄调查的可变费用比例更大。

尽管认识到了这些问题，却极少有针对研究费用的研究报告，因而我们只能做一般性的了解。面访一般比电访要贵很多。曾有报告说，这两者的费用之比大约为2，也就是说，每一个面访的调查费用是电访的2倍（见Warner，Berman，Weyant，& Ciarlo，1983；Weeks，

Kulka， Lessler， & Whitmore， 1983；Van Campen， Sixma，Kerssens，& Peters，1998）。不过，根据我们的经验，对全国性调查而言，这个比率要更高，5到10倍都有可能。在面访的费用中，很多都花在了访员到调查对象所在位置的路上了，而每次重复拨打电话却只是略微增加调查的费用。此外，因为调查地点分散，实地访员通常需要比电话访员需要更多的和更好的训练，以使其具备更多的经验；对电话访员而言，任务要简单许多，他们集中工作，并且可以集中监管。如果使用计算机辅助访问，面访与电访之间的费用差距会增大，因为还有额外设备的费用。在实地访谈中，必须要为每位访员配备笔记本电脑，且其使用率不及集中电访使用的台式电脑。另一方面，用于CAI设备的固定成本不受调查模式的影响，因此整体费用的差异往往还是依赖于样本量的大小。

电访模式与邮寄模式的费用差异一般比较小，其比率大概在1.2倍（如，Hochstim，1967；Walker & Restuccia，1984）到1.7倍（如，McHorney， Kosinski， & Ware， 1994；O'Toole， Battistutta，Long， & Crouch， 1986；Warner， Berman， Weyant， & Ciarlo，1983）之间。另外，这也依赖于寄出或回收的数量，以及其他许多因素（如样本量大小、计算机的使用等）。

尽管很多人都说，互联网调查比邮寄调查要便宜很多，但这两种方法费用的比较还是要看预算中包括了哪些，以及工作量的大小。网络调查的固定费用一般要高于邮寄调查，主要是基础设备的费用，还有用来开发、测试特定调查问卷的费用。另一方面，如果整个调查都用电子版本（例如，使用电子邮件来邀请和提醒），那么平均到每一份互联网调查的费用，基本接近于零。相反，邮寄调查的固定费用通常很小，可变费用（打印、邮资、电报或扫描等）却比互联网调查大

很多。因此，这两种方法的费用比较，很大程度上取决于样本的数量，要看把固定费用平摊到每位受访者身上会有多少。

显然，具体到某个调查的费用结构，还涉及许多其他因素。例如，面访调查中加入语音-CASI部分，就会比单一访员指导下的费用增多，当然，也会提高数据质量。同样，在电话调查中加入邮寄的广告（为了提高合法性，也可以表达诚意）也会增加费用，这样，也能提高应答率。我们应该清楚，对不同模式费用比较的泛泛概括，要谨慎处理，像调查质量的其他因素一样，对费用，也要把各种设计要素综合起来一起考量。

可以想象一下，如果将来的调查设计涵盖多种数据搜集方法，就需要对不同方法的费用模式开展研究。如果调查研究者希望以最佳方式利用好资源（如在有限的预算下使统计质量最优），那么，对不同方法的误差及其相对费用，就都需要好好研究。

5.3.7　方法选择小结

对于一个跨区域的新调查而言，数据搜集方法的选择，需要权衡不同误差来源对整体估计的相对影响，以及其他因素，比如费用、时间、人员和其他资源的可获得性，等等。由于费用昂贵，面访通常只用于由联邦政府支持的大范围、目标群体覆盖率高、应答率高、数据质量水平高的调查。我们给出的三个调查范例（NCVS，NSDUH，NAEP），基本上都靠面访的数据搜集。为了获得要求的覆盖率、应答比率和数据质量，至少对受访者的首次访谈要使用面访模式。其余还

有两个调查范例（SOC，BRFSS）使用了电话模式。CES则是典型的混合型模式。

对随时间发展持续搜集数据的调查而言，更换调查方法会把情况变得更为复杂。在考虑新方法优势的同时，也需要考虑加入新模式会带来的不稳定因素。因此，从一种方法转换到另一种方法时，一定要谨慎；通常，还需要通过对分样本来实验，来判断由转换所带来的影响。对长期或跟踪调查而言，如果要转换调查方式，就会面对同样的问题。从单一数据搜集模式变为多种模式（见5.4节 ）也要留意这个问题。我们前面讨论的许多模式比较研究，都是为了了解模式转换的影响。

一般而言，对各种数据搜集模式的测量误差大家都已经相当清楚了。可比的等价模式（如面访与电访；邮寄与互联网），结果的估计常常差别不大。唯一的例外是对高度敏感行为的测量，如药物滥用，性行为之类（见图5.4）。在这类测量中，自访模式的优势确凿无疑。对大量非敏感性访题而言，访员指导的调查，无论是电访还是面访，得到的结果都很相似；就自访答而言，无论是在纸版还是网上，结果也都差不多。因此，需要衡量且要区别对待的，是非测量误差（覆盖和无应答）和效率（时间、费用等）等因素。

5.4　运用多种数据搜集模式

许多调查都在使用多种数据搜集模式。有许多理由促使一个调查用一种以上的模式，有三条尤其常见。第一，使用混合模式可以降低费用。一般说来，首先试着用最便宜的模式（邮寄）从每个受访者那里收集数据，然后换稍贵一点的模式（电话）去收集那些在最初邮寄

模式中没有应答的数据，最后，再对剩下的样本和个案运用面访。在这种设计中，越是花钱的模式，调查的样本就越少。美国2000年的人口普查就运用了这种设计，开始使用邮寄问卷，然后改为面访，跟进那些无应答的受访者。（在一些场合下，也会使用CATI，IVR和互联网数据搜集方法。）

使用混合模式的第二个理由是要使应答率最大化。如既有的大型调查“当前就业统计调查”（CES）就采用了多种数据搜集方法，包括互联网、传真、语音电话数据输入（IVR输入）、电访和邮寄。其中，对无应答（时间有限也很重要）的考虑超出了对模式效应的考虑。提供混合模式的方式是，允许受访者选择对他们来说最合适的方法来参与调查。

用混合模式的第三个理由是为了在历时调查中节约成本。我们提过，在CPS和NCVS的第一轮调查中使用了面访，即让应答率最大化（通常在长期调查的第一轮，无应答的损失最严重），访员也可以趁机获得住户的电话号码，为后面的调查打开方便之门。如此。后面的调查就可以通过电话来收集了，进而也节约了成本。

不过，上面提到的，只是混合模式的一些可能。对模式的选择取决于是关注受访者（如对一些受访者用面访，其余的用电访；或一些用邮寄，另一些用互联网），还是关注某个阶段（如用电话作招募，用IVR完成调查；或一开始用邮寄，用电话作为提醒），甚或关注某类访题（如一些访题用语音-CASI，另外一些由访员提问）。图5.5列出了混合模式方法的几种形式（也见de Leeuw，2005）。

1.对部分受访者使用一种模式,对另一部分受访者使用另一种模式

- 如电话调查,对没有电话的受访者使用面访
- 如邮寄调查,配合互联网调查

2.用一种模式来招募受访者,用另一种模式来调查

- 如用邮件邀请受访者参加互联网调查
- 如用电话为 IVR 调查招募受访者

3.用一种模式搜集数据,用另一种模式提醒或跟进

- 如电话提醒参加邮寄或互联网调查

4.在历时调查的第一轮使用一种模式,后面的几轮使用另外的模式

- 如第一轮用面访,后面几轮用电访或邮寄

5.调查的主要部分使用一种模式,少数项目使用另外一种模式

- 如用语音-CASI 来访问敏感项目

图5.5　五种不同类型的混合模式设计

混合模式的数据搜集方法之所以会崛起，还要归究于上面讨论过的大量研究发现。这些发现表明，不同的模式得到的数据结果一般差不多，但在覆盖率、无应答、费用等方面常会有区别。举个例子，一旦得知电访与面访调查在测量误差上效果相仿，在很多情况下，研究者就会替换掉面访调查（如模式更换）和补充电访（混合模式设计）。同样，也可以用互联网调查来补充单一的邮寄问卷调查。混合模式设计的逻辑是，通过方法的组合，既可以发挥每种模式的优势（如电访可以节约成本），还可以弥补各自的不足（如覆盖率）。因此，混合模式设计牵涉到对误差来源的具体测量。

明确两种差异是非常关键的，即用同一种模式获得的数据，与同样一套访题用不同模式获得的数据之间的差异；以及因不同受访群体而造成的差异。处理用不同方法且又来自不同样本群体的数据时，关键要确保无论样本有何不同的特征都不存在任何的模式效应。尤其是在受访者选择自己喜欢的方式，或因条件限制被迫只能使用一种模式时，更要关注这两种差异。研究者可以在调查中或调查前就认真推敲

设计中的模式比较（如在电访与面访案例），或也可以将模式比较放进访题中，把少量受访者随机放入不同的模式来作比较，这样就可以帮助分辨上面的两类数据差异，明确哪些反映了总体的真正差异，哪些是来自不同数据搜集模式的差异。

混合模式调查设计要考虑的问题与单一模式的不同。在混合设计中，不是要做到尽量优化某种模式，而要运用工具与程序，实现模式之间的匹配。例如，设计一个互联网调查工具，就要尽可能做到与纸版调查相近，并可以运用互联网调查的功能避免复杂的跳转、随机化、编辑检查之类。同样，如果是有电访在内的混合模式设计，还必须考虑在面访时要不要使用“选项卡”，目的是让不同模式的基本测量条件能达到尽量的相似。

混合模式设计也加大了数据搜集的复杂性。要避免重复，即避免对同一位受访者用两种模式收集两次数据，因此样本管理就变得更加重要。从一种模式转到另一种模式的时间点，也必须仔细考虑。假如不同模式会产生不同类型的误差，尤其当自访和访员指导的模式混合时，在做数据清理和汇集时，必须要留意是不是只生成了一个分析文件。总之，混合模式的设计越来越受欢迎，这其中有很多原因，最主要是可以节省费用，加快数据搜集的速度并提高应答率。为了追求这些目标，我们期待不久的将来会有更多创造性的混合模式设计。

5.5　小结

在过去的几十年里，邮件、电访和面访方法是调查方法最主要的三种模式，现在则出现了多样化的数据搜集方法。这些方法的不同之处在于是否需要访员或要访员做什么，在多大程度上要与样本直接接

触，测量中可以为受访者提供的隐私保护程度如何，使用的沟通渠道是什么，以及是否或如何使用计算机辅助技术。这些特点对调查结果的统计误差都会有影响。

选择调查方法需要针对具体的方法考虑众多因素，比如抽样框的覆盖率和接触率；针对研究主题，方法是否恰当；调查费用的约束；以及结果的即时价值等。

调查方法包含了对数据搜集方法效果研究的许多随机实验。有些实验为了测量沟通渠道的纯效果；有些则比较不同设计的典型特点。面访和电访模式常用的抽样框，与邮件和电子邮件的抽样框相比，对家户总体有更好的覆盖率。就应答率而言，典型方法的排序依次是：面访、电访、邮寄，最后是互联网调查。这些结果，大部分都基于对成年目标总体的家户调查；如果是其他总体，排序可能会有不同。就提高应答率而言，使用访员看上去是最有效的。由于社会期许效应，自访模式的应答误差会比较低。访员指导下的选项缺失率往往很低。在选择封闭式访题的答案时，如果采用语音模式来呈现访题，一般会出现近因效应；如果使用视觉呈现，应答容易产生首因效应。

不同的数据搜集方法需要的费用有很大差别。一般而言，从最高到最低的排序是：面访、电访、邮寄和互联网调查。费用比较对样本规模很敏感，不同方法在固定费用（与样本大小无关）与可变费用（随样本大小同方向变化）上的相对比率也各不相同。

混合模式设计的应用越来越广，因为可以更好地在费用与调查结果误差之间进行平衡。历时调查经常在开始时使用面访，在后面的调查中换成便宜一些的方法。

随着既省钱又可提高数据质量的新技术的出现，会不断发展出更多的模式。因此，理解方法之间的不同，以及这些不同背后的特点或维度，也就变得越来越重要。只有充分理解这些特征，才能在具体研究中选择使用哪种或哪几种方法，进而作出好的决策，才能评价和采纳那些最新出现的、目前还没有得到检验的模式；才能创造性地综合不同方法的优势，实现调查费用和误差的最小化。

关键词

计算机辅助语音自访（ACASI）

计算机辅助个访（CAPI）

计算机辅助电话访问（CATI）

计算机辅助自访问卷（CSAQ）

极端效应（extremness）

互动式语音应答（IVR）

开放题（open-ended questions）

近呈效应（recency effects）

自访问卷（SAQ）

社会期许偏差（social desirability bias）

电话-语音计算机辅助自访（T-ACASI）

电话键盘数据录入（TDE）

默许效应（acquiescence）

计算机辅助自访（CASI）

沟通渠道（channels of communication）

磁盘邮寄方法（disk by mail）

固定费用（fixed costs）

模式（mode）

首呈效应（primacy effects）

应答效应（response effects）

选项卡（show cards）

社会在场（social presence）

文字访题和应答记录（text-CASI）

互联网调查（web survey）

进一步阅读资料

Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls, W., and O'Reilly, J. (1998), *Computer*

Assisted Survey Information Collection , New York: Wiley.

Dillman, D., Smyth, J., and Christian, L. (2009), *Internet , Mail , and Mixed Mode Surveys : The Tailored Design Method* , 3rd Edition, New York: Wiley.

Gwartney, P. (2007), *The Telephone Interviewer's Handbook : How to Conduct Standardized Conversations* , New York: Wiley.

Lepkowski, J., Tucker, C., Brick, J.M., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (2008), *Advances in Telephone Survey Methodology* , New York: Wiley.

作业

---

1. 一个非营利公益组织“智库”计划开展一项调查，了解国内低收入家庭的状况。问卷访题有许多内容，包括家庭收入、国家福利项目的参与、低收入家庭的健康保险福利、子女教育和健康状况、父母的就业情况等。研究者对农村及内陆城市的低收入家庭尤其感兴趣。研究者考虑使用两种模式：CATI电话调查或者CAPI面访。讨论该项目采用这两种模式各自的好处与缺点。

2. 简要说明使用访谈法收集调查数据的两种好处与两种不足。

3. 如果在面访中使用自访答问卷法搜集性行为/经历的信息，请指出一个好处与不足。

4. 下面哪种数据搜集模式最不容易产生访题的顺序效应：面访、电访、邮寄方式？说明理由。

5. 针对下面提到的每一条要求，请说明在电访、面访和邮寄三种方式中，哪种最适宜于家户调查，并说明原因。

（a）最快完成调查。

（b）在样本大小一致的条件下，费用最少。

（c）应答率最高。

（d）如果想除了测量一般人以外，还能很好测量到少数说不同语言的人群。

（e）在样本大小一致的条件下，抽样误差最小。

6. 你是否同意下面的每种说法，说出自己的理由。

（a）用CAPI进行美国家户调查，覆盖误差是主要的顾虑。

（b）用互联网进行美国家户调查，覆盖误差是主要的顾虑。

（c）互联网调查的受访者数量很大，说明可以得到更好的调查估计（如，对总体的估计）。

（d）如果使用链接进入的方式进行互联网调查，对评估无应答率会更容易。

7. 你所在的市场研究公司要进行一项互联网调查，以估计成年消费者对一种新饮品的购买可能。客户有足够的经费，但要求在两周内

提供估计值与“误差区间”。你的老板不太确定是否要接下这单委托，请你来做可行性分析，并要逐一列出用互联网做这类调查的优势和劣势。请列出这种数据搜集方法的两个好处与两个不足，并且给你的老板建议是否接受这项委托。

8. 就上述情形，如果不采用互联网调查，客户想使用邮寄邀请加回拨语音调查，怎样呢？举出IVR搜集数据的两种局限。

9. 你所在的地方政府请你设计一项调查来评估当地居民对公共交通的看法，并评价他们希望更换交通模式的程度。你会建议他们使用哪一种数据搜集模式，为什么？

10. 美国统计学会希望对他们的美国会员做一项调查，了解他们对待统计学工作者工作资质的看法。你会建议使用哪种数据搜集模式？请给出你的理由。

11. 比较电访与面访两种方法得到的统计数据，思考由于沟通媒介的不同，并列出这两种设计之间的三种差异。

12. 与电访相比，家户面访为什么经常会得到更好的应答率？请给出两个理由。

13. 有一项调查设计，要研究不同收入水平以及拥有不同健康保险的人们健康照料的可及性。在研究的进程中，资助方突然缩减了25%的经费。考虑到经费的约束，客户建议将研究由基于区域概率抽样的家户面访改为随机拨号的电访。依照客户的建议，他们希望可以保持样本规模。如果采纳客户的意见，必须考虑哪两类误差来源？举例子说明为什么改变设计会影响这两类错误来源。

14. 如果是混合设计，先使用邮寄调查，然后对无应答的受访者使用电访，再对仍旧无应答的受访者使用面访调查，请举出这种混合模式设计受欢迎和不受欢迎的两个方面。

15. 如果是混合设计，先使用邮寄调查，然后对无应答受访者使用电访，再对仍旧无应答的受访者使用面访调查，请列出一个优势和一个劣势。

16. 如果要进行一项全国性调查，请简要列举出影响面访和电访成本差异的因素。

17. 请简要说明在电访调查中，只使用手机的用户给误差带来的影响。

（a）覆盖误差

（b）无应答误差

（c）测量误差

18. 简要说明CASI和CSAQ的区别。

19. 简要说明呼入与呼出语音互动式语音应答调查的区别以及两类调查的误差。

20. 简要说明集中电访与分散电访的优势。

6　抽样调查中的无应答

6.1　导言

第1章和第2章讨论过调查研究者如何运用“无应答”（nonresponse）概念描述无法获得对样本单位测量的情形。有时是彻底无法获得测量，被抽中的样本完全拒绝合作并接受调查（例如受访者说：“我从来不参加调查，请别再打电话来了。”）；有时，只是某访题无法获得测量（如访员：“去年您的家庭总收入是多少？”受访者：“我不知道，只有我妻子清楚这些事”）。完全无法获得测量的情况，被称为“样本无回答”（unit nonreponse），又称为“单位无应答”，部分无法获得测量的情况，被称为“访题无应答”（item nonresponse）。

无应答会影响调查统计量的质量。如果无应答的受访者在某个变量上的取值与其他人取值的平均水平相异，而这个变量又是调查估计值计算的一部分，那么仅基于应答计算的统计量会与理论上基于所有样本计算的统计量有差异。如果差异是系统性的，差异大小随调查设计不同而不同，那么这就是“无应答偏差”（nonresponse bias）。某些简单的统计量（如样本均值），其无应答偏差是无应答样本占总样本的比例以及应答与无应答样本在该统计量上取值差异的函数（见2.3.6节）。

在美国和欧洲的大多数家户抽样调查中，无应答率（即无应答样本数占合格总样本数的百分比）正逐年升高。在各种抽样调查方法中，无应答主要有三个主要来源：无法与被抽中的样本取得联系并发出调查邀请；虽向样本发出了调查邀请，却无法得到其同意和配合；受访者没有能力提供调查需要的信息。不同的无应答来源会影响不同类型的调查统计量。

本章将向读者介绍与无应答相关的基本概念和实践。首先介绍应答率及其发展趋势，接着讨论无应答率与偏差之间的关联，然后解剖不同类型的无应答现象。在结束的时候，讨论调查设计特征如何影响由无应答带来的统计量误差。

6.2　应答率

简单地说，一项调查的应答率就是合格样本应答的百分比。无应答率就是应答率百分比的补值。在讨论应答率的特征值之前，我们需要注意到这个简单的定义实际上包含了对应答率与无应答率的复杂计算。

6.2.1　计算应答率

因为，传统上，无应答率被作为测量质量的一个直接指标，对高质量的统计量而言，无应答显然是令人沮丧的。这也是许多专业组织的焦点问题（Frankel，1983；American Association for Public Opinion Research，2000）。我们可以在美国舆论研究协会（the

American Association for Public Opinion Research，AAPOR）的网站（www.appor.org）上找到一些手册，根据研究设计，手册中提供了多种不同的应答率计算方法。一般而言，在计算应答率时，有三类复杂的情形：

1）某些抽样框的单元并非目标总体成员，需要通过过滤来确认其合格性（例如，针对家户的电访调查如果有商用电话）。在这样的设计中，就很难确认在无应答受访者中的合格性，进而也很难确定应答率的分母是什么。

2）某些抽样框包含有群样本要素，其中，在抽样时，要素的量是未知的（例如，从一所学校中抽取儿童样本）。当整个群无应答时，就很难确认有多少要素是无应答者。

3）在抽样框中，不同的要素有不同的备选概率（例如，针对少数族群的过度抽样）。在这种情况下，就不知道在计算应答率时是否需要加权。（参见Groves，1989）

解决第1）、2）个难题的方法之一，就是估计分母的值，要么运用外部信息，要么运用其他例子的信息。故可能的应答率为

$$\frac{I}{I+R+NC+O+e(UH+UO)}$$

式中　I ——完成了的访问；

R ——拒访和中断的访问；

NC ——未联系上；

O ——其他合格样本；

UH ——未知是否为家户；

UO ——未知是否为合格样本；

e ——未知是否合格样本被作为合格样本的估计比例。

对e 估计，可以从当下的调查中获得，例如（$I+R+NC+O$）/（$I+R+NC+O+$不合格但被抽中的样本）。此外，也可以用一些特殊的方法来计算e，例如研究一群最初并不知道其是否合格的样本。最后，如果不能获得e，就建议报告两类应答率：其中一类在分母中包括（$UH+UO$），另一类则不包括。如此，就产生了一个应答率的区间，其中一定有真实的应答率。在Brick，Montaquila，和Scheuren（2002）的研究中，还呈现了复杂的、通过建模计算e 的方法。

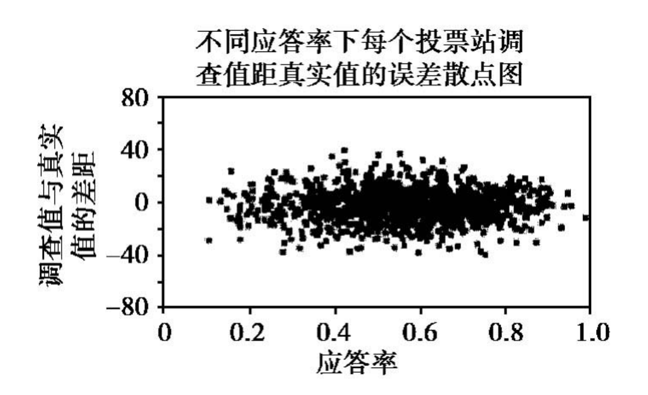
Merkle and Edelman（2002）关于无应答率如何影响无应答误差的讨论

Merkle和Edelman（2002）呈现了一个观察调查，说明无应答率和无应答误差之间没有关系。

研究设计：对结束投票正准备离开投票站的民众进行调查。对投票地点采取概率抽样，对投票者采用系统抽样。调查每一个样本

投票站，受访者回答投民主党和投共和党的人数之差，把这个结果与公布的总体投票结果进行对比。

研究发现：各调查现场的应答率从10%～90%不等，大多数位于45%～75%。下面的散点图中，每一个点代表一个投票站的数据；*X* 轴代表调查现场的应答率；*Y* 轴是民主党和共和党得票比例差的误差。应答率和误差之间没有显著的相关性。对于受访者是否合作的一个好的评价指标是访员站立的位置距离出口的远近。



（数据来源：Merkle and Edelman，2002）

研究局限：由于对投票站的访员安排未做随机化处理，可能造成访员效应和真实回答率之间的混淆。而且，投票站是一个特殊场景，不能完全代表其他类型的社会调查的情况。

研究意义：提供了一个实例，当造成无应答的原因与统计量不相关时，不会造成无应答误差。

如果在抽样中出现不等概率（正如第4章讨论的），就会出现第三个难题。例如，如果在城市服务的社区调查中针对穷人社区（第一层）过度抽样，其抽样比是其他区域（第二层）的两倍，如何计算应

答率呢？在这样的设计中，通常有两类应答率问题。第一是第一层与第二层应答率的比较，与两个层的均值比较相关。在这种情况下，对两个应答率的使用宜采用相同的方式。如果要计算总的样本均值，正如第4.5节说明的，就必须使用权重w_i，然后，针对总的应答率，也要使用相同的权重，即对每个样本要素的入选概率进行调整。

在不同的情况下，可以使用不同的率。例如，拒访率（$R/(I+R)$）和拒访转换率（初始为拒访，后来又访到的比率），用来评估访员的绩效。对既有的调查，就会用到覆盖率（访到的占应访单元的比例）；如果要估计产出或雇员数量，丢掉了沃尔玛与丢掉了便利店可不是同一回事儿。类似于全国教育进展调查（NAEP），在多个层级（如学校、学生）都有选择，每一层都要计算无应答，进而建构一个综合率。

6.2.2　应答率的历时变化趋势

应答率的变异性可以借用正在进行的调查其应答率的历时变化进行探讨。例如，图6.1是全国刑事犯罪受害者调查（NCVS）历年的应答率数据。NCVS是对样本家庭所有12岁及以上家庭成员进行的调查。NCVS报告了家庭和个人的两类应答率。家庭的无应答率指所有家庭成员都没被访到的家庭所占的百分比，如图6.1上方的虚线所示。这个值，历年的数据比较稳定，变化幅度在3～4个百分点（家庭的应答率约为96%～97%）。最下面的那条线（灰线）是家庭的拒访率，这个值历年来似乎也维持在相对稳定的水平。图中黑色的实线，是个体的拒访率，即调查名单上列出的全部应该访问到的家庭成员中拒绝接受调查的家庭成员所占的比例。简单地说，就是在与样本家庭取得联系的

前提下，大约有百分之多少的受访者同意并接受了访问。从图中可以看到，在图中所涉及的时间段内，该项无应答指标值增长了近2倍。

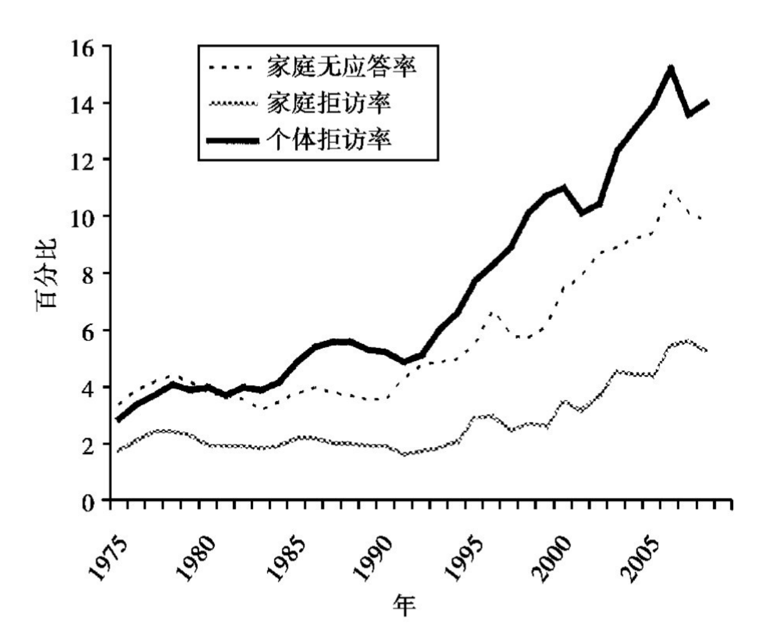


图6.1　全国刑事犯罪受害者调查中历年家庭无应答率、家庭拒访率和个体拒访率曲线图（数据来源：U.S. Census Bureau，2007）

图6.2展示的是当前就业统计调查（CES）历年应答率的变化，这是一项由美国劳工统计局资助，由美国人口普查局组织实施的调查，根据调查结果每月定期发布美国当月的失业率数据。图中数据的变化趋势与上例完全不同：无应答率数据在大多数年份中总体稳定在4%～5%，而拒访率却呈现持续上升的趋势。在面对高拒绝率压力的情况下，这是一个通过努力降低无接触率（noncontact rate）使总无应答率维持在较低水平的一个实例。那么，是什么造成了1994年无应答率的大幅上升？原因是那一年的调查方案有较大调整，改成了计算机辅助面访，并且采用了新问卷。调查设计的改变之所以会给应答率造成如此影响，可能的解释是：纸版的简短问卷（8～12分钟就能完成）往往让受访者站在自家的台阶上就能回答完，采用笔记本电脑为辅助

工具进行调查的问卷就不能以这种方式方便地进行回答，这样就造成了拒绝率的升高（Couper，1996）。

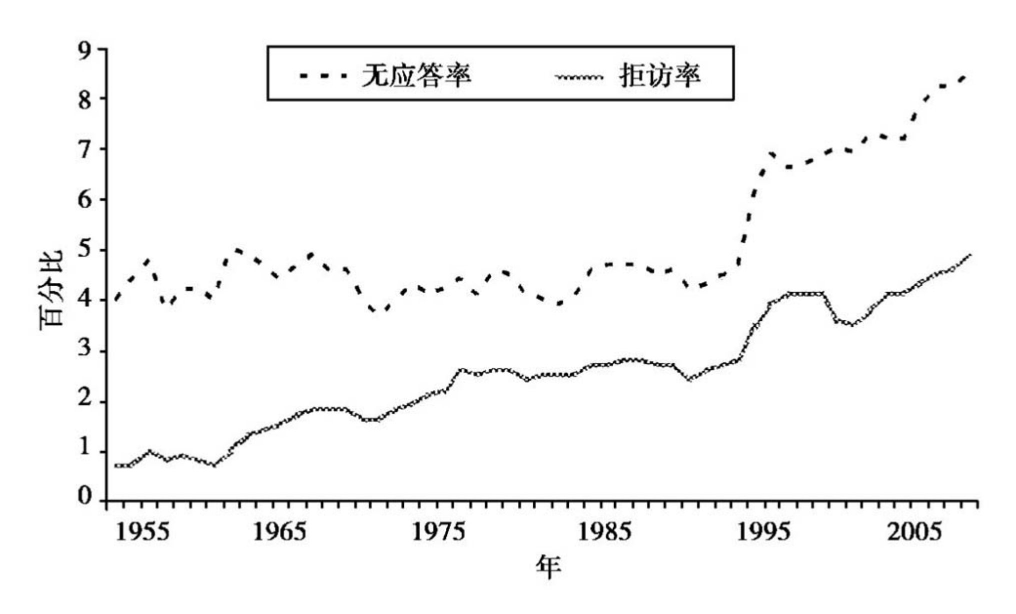


图6.2　当前就业统计调查历年无应答率及拒访率图（数据来源：U.S. Census Bureau）

上述两项调查的无应答率都很低，不过，总地来说，无应答率普遍存在上升的趋势。这两项调查都有充分的资助，也是以面访形式进行的调查，并且都是在联邦政府主持下开展的。一般而言，与之相比，其他类型调查的应答率更低，学术性调查的应答率比这两个还要低一些，私人部门组织调查的应答率就更低了。

图6.3是消费者调查（SOC）的总体无应答率和拒访率趋势。这项调查的无应答率要远远高于NCVS和CPS的。总体无应答率呈现稳定的上升趋势，从1980年代的30%左右增长到近些年的将近60%。拒访率曲线也呈现相类似的上升趋势。同时，先前拒绝后来接受了调查的这类样本的比例从7%增长到了15%。这又一次证明：要是没有访员团队在说服调查者参与方面付出了艰苦的努力，调查应答率将会更低。

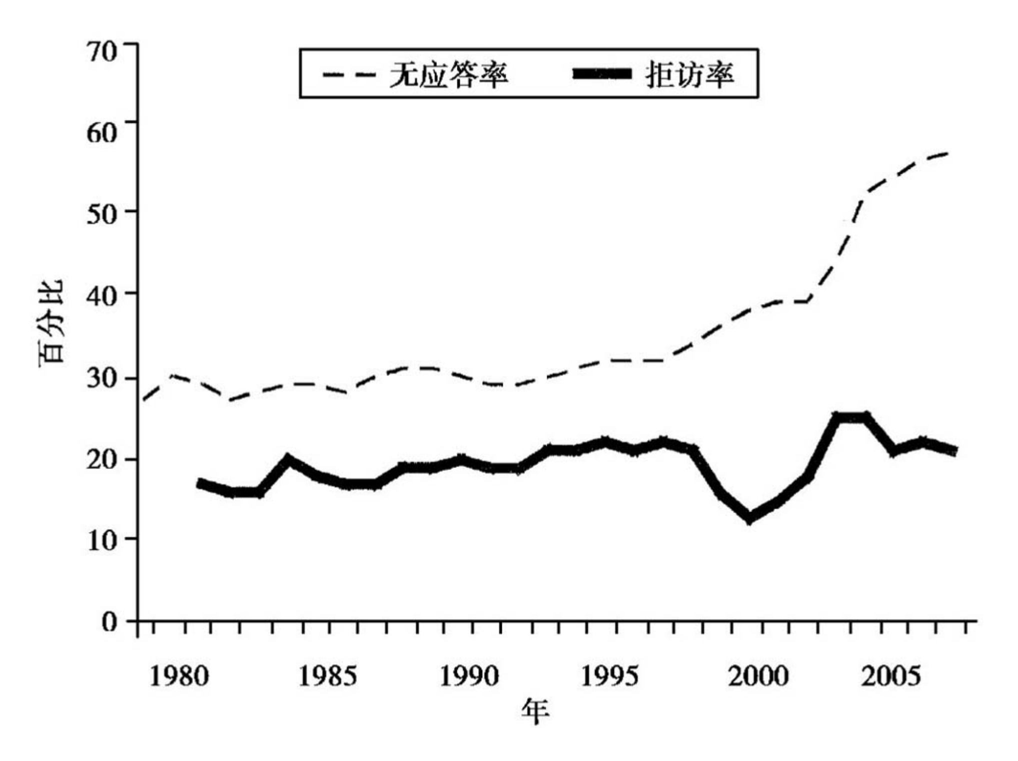


图6.3　消费者调查历年无应答率及拒访率图（数据来源：Survey of Consumers）

图6.4是行为风险因素监测系统（BRFSS）历年的无应答率数据。由于这项调查要在每个州分别计算个体无应答率，因此，图中的无应答率数据是所有州无应答率的中位数。从图中可以看出，这项调查的无应答率，更类似于SOC，而不是更接近NCVS。这大概是因为，首先BRFSS和SOC均采用了电访，而NCVS是面访。其次，美国政府只是BRFSS的间接资助方。BRFSS的总体无应答率在这些年中，从开始的30%多增长到了后来的约50%。

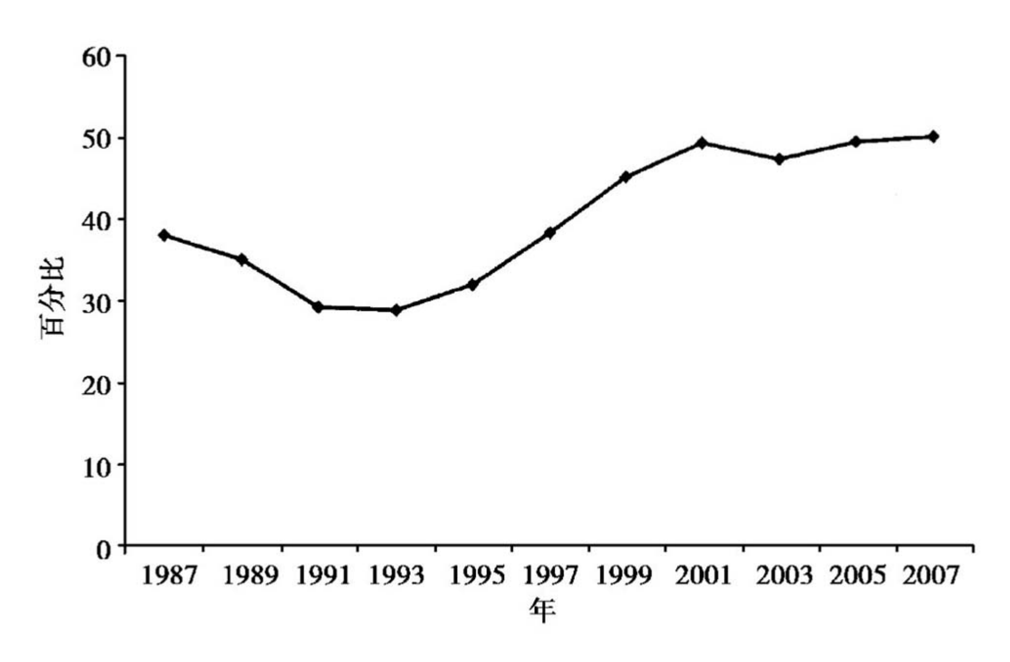


图6.4　行为风险因素监测系统历年中位数无应答率（数据来源：BRFSS）

无应答率的持续增长不仅仅出现在美国。一项针对欧洲16个国家直到20世纪90年代20多年的调查发现，与调查样本无法取得接触的比例每年平均增长0.2%，拒访率每年增长0.3%（de Leeuw and Heer，2002）。

以上都是家户调查的结果。商业机构进行的调查经常采用邮寄或电访（或二者结合）方式。在商业机构的调查中，较少见到无应答率增长的证据（Atrostic and Burt，1999）。

6.3　无应答对调查估计值质量的影响

无应答率对调查质量影响的观点变化较多。在前几十年里，调查研究者的主要目的就是让无应答率最小化。例如Alreck和

Settle（1995，p. 184）说，“尽量降低无应答率显然是重要的，因此要鼓励有足够的应答率。”Babbie（2004）更加武断地说，“我相信应答率至少要有50%才可以用于分析和报告。60%算是好的，70%就很好了。”（此处 ）最后，Singleton和Straits（2005）提到，“因此，关注应答率非常重要。对问卷调查而言，应答率最低要有80%；低于70%就很容易产生偏差。”（此处 ）所有这些，都引自调查方法的教科书。

把降低无应答率作为降低无应答误差的唯一方法，来自于这个公式：

$$\overline{Y}_{r} = \overline{Y}_{n} + \left(\frac{M}{N}\right)(\overline{Y}_{r} - \overline{Y}_{m})$$

这里$\overline{Y}_{r}$ 是未经调整的应答均值，$\overline{Y}_{n}$ 是全样本均值，M/N 是无应答者所占的比例，$\overline{Y}_{m}$ 是无应答均值（在大多数调查中，是一个未知数）。许多研究者探讨的是双变量情形下应答均值误差，认为（$\overline{Y}_{r}-\overline{Y}_{m}$ ）是一个固定值，进而认为减少*无应答误差* （nonresponse bias）的唯一途径就是降低无应答率。

上面的公式在针对调查无应答时，又被叫作“决定系数”。也就是说，在全样本中有一个固定的应答者集和无应答者集。对所有样本而言，调查让研究者看到的是，谁是应答者，谁是无应答者。与这个观点不同，更加现代的观点认为，每一个个体都是潜在的应答者和潜在的无应答者，最终是否参与是一个随机过程（即一个随即变异过程）。在这个观点下，上面的公式就可以表述为：

$$\overline{Y}_r = \overline{Y}_m + \frac{\sigma_{yp}}{\overline{p}}$$

式中，σ_{yp} 是方差，在所有样本单元中，y 是调查所针对的变量，p 为应答倾向性；$\overline{p}$ 样本单元中应答倾向性的均值。方差测量的是两个变量的共变性。

$$\sum_{i=1}^{N}(y_i - \overline{y})(p_i - \overline{p})/(N-1)$$

上面的公式意味着，应答倾向性 （response propensity）与其感兴趣的变量之间有着强相关。因此，对应答均值而言，无应答误差应该大。简而言之，在一个调查中，针对不同的统计量，作为y 和p 之间方差的函数，无应答误差应该是不同的。除了关注应答率之外，研究者必须关注调查变量及其关联的应答倾向。

是否有经验数据支持说在一项调查中不同的估计值之间无应答误差的变异很大？有的。一项元分析整合了针对同一现象的需要科学研究的发现。有一类调查，针对的就是调查估计值的无应答误差。有时候，这类对应答者和无应答者之间在某些变量上的比较已经包含在抽样框中；有时候，包含在前期的过滤性调查中；有时候，则包含在后来对无应答者的随访中。我们注意到，在所有估计值中，无应答率从14%到72%，无应答率的均值为36%。大多数的估计值应用了非调查记录（24%来自抽样框，32%来自于补充数据集）；28%应用了随后对无应答者的极端努力。还有一些应用了过滤性调查数据（14%）。极少数（2%）提出，在未来的调查中要注意。

图6.5展示了959个相对无应答偏差的绝对值的估计值。

$$\left|\frac{100\times(\bar{y}_r-\bar{y}_n)}{y_n}\right|$$

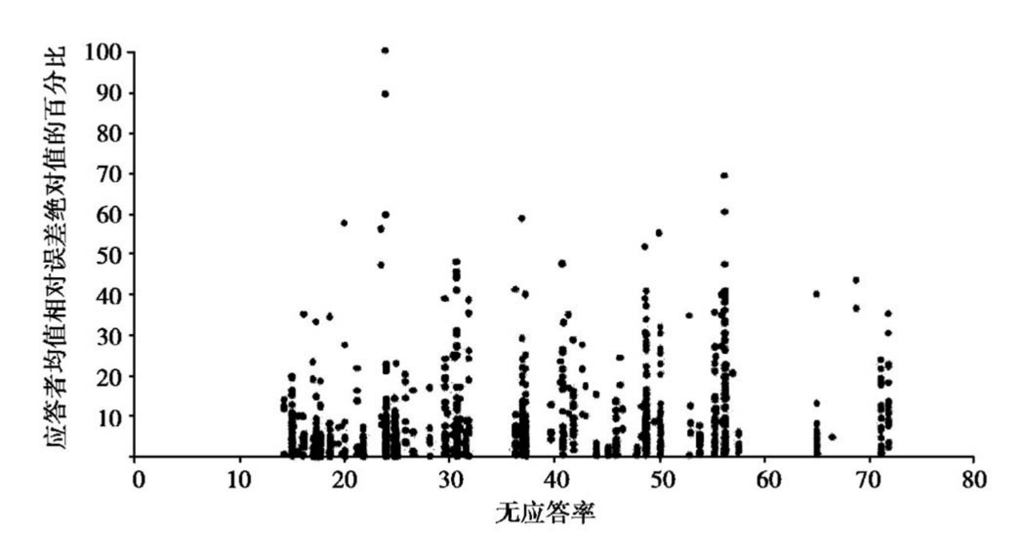


图6.5　959个应答者均值相对误差绝对值与无应答率的统计量（数据来源：Groves and Peytcheva，2008）

这里分子中包含了应答者均值与全样本均值之间的差，分母则是全样本均值。图中包含了59项研究中每一个均值的点，以及双变量之间的互补百分比。对每一个双变量，可以计算两个百分比。两个百分比中，小的那个可能会产生较大的相对无应答误差。因此，图中展示了两个互补百分比的无应答误差，显示为纵向顺序的点，代表的是在同一项调查中计算的不同估计值。这幅图明显显示，（a）在研究中有大的无应答误差；（b）在同一项调查中，无应答的大多数变异，正如观察到的，都与估计值有关；（c）调查无应答率本身，对相对无应答误差的绝对值而言，是一个较弱的预测变量。如果拟合一条线性回归

线，则R^2会等于0.04。简而言之，把无应答率和无应答误差关联起来需要调查中每一个测量的更多的环境信息。

图6.5对调查研究者而言，至少有两项应用。首先，经常会遇到极小的σ_{yp}，即感兴趣的变量与应答倾向之间无关，因此，无应答误差会很小（无论调查的应答率是多少）。第二，发现应答倾向如何与重要的调查变量有关是研究者试图降低无应答误差的一项新任务。这需要在因果模式下来探讨无应答误差。关于上面的图，有一点需要提醒。由于点阵图来自于不同的调查，故没有对此提供答案。“对某个具体调查而言，提高应答率会降低无应答误差吗？”有时候，回答会是“是”，有时候则是“否”。不过，很少有这样的例子，在增加应答率的同时也增加无应答误差（参见Merkle，Edelman，Dykeman，and Brogan，1998）。答案取决于不同类型的无应答者是否被带进了应答者群，即应答率的提高是增大了还是缩小了σ_{yp}。

非常重要的是，还要注意到，无应答可能会影响到描述性统计和分析性统计（例如回归系数）。无应答误差对这些估计值影响的表述，比上述的议题更加复杂，通常是相关变量方差的函数。已经有大量的建模技术可以用来在这些统计量中讨论无应答误差，大多数都在计量经济学领域（Heckman，1979；Berk，1983）。

下一节，我们会讨论，当无应答率对某个统计量的无应答误差在我们的理解中变得非常复杂时，就需要考虑无应答的因果机制了。

6.4　调查无应答误差的因果思考

应答倾向性与无应答误差之间的关系，有三种可能的因果模式（Groves，2006）。正如图6.6所示，“分因”模型认为，Y 变量的原因独立于应答倾向P 的原因。在这种情况下，依据应答者对Y 的期望值就是对全样本的无偏估计以及对“完全随机缺损”的对应（Rubin，1987）。“共因”模型认为，在应答倾向和变量Y 之间，有着共同的原因（Z ），这个模型对应随机缺损情形。“调查变量因”模型认为Y 本身就是调查倾向的原因，是不可忽视无应答 （nonignorable nonresponse）的条件。

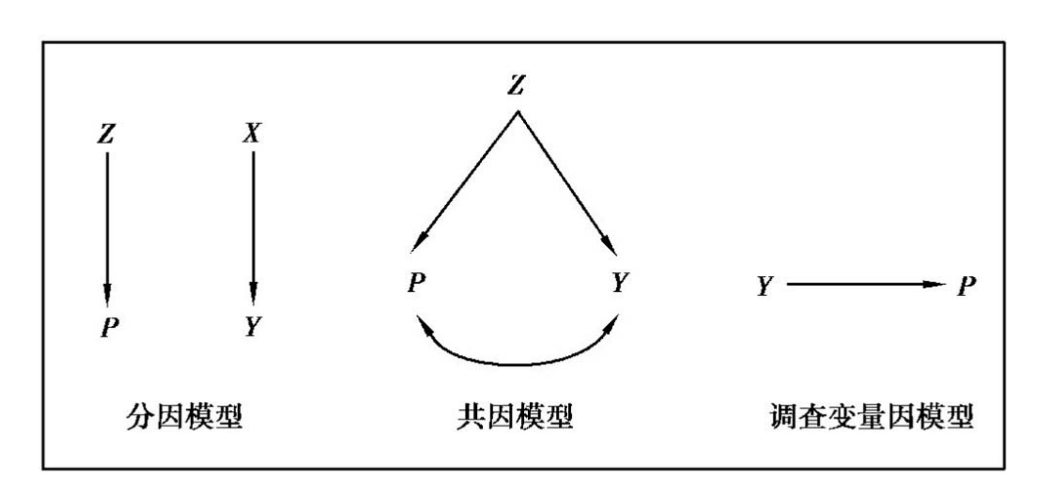


图6.6 应答倾向（P ）与调查变量（Y ）的方差，包括辅助变量（S ，Z ）（数据来源：Groves，2006）

在上述的每一个无应答误差模型中，简单应答平均数可以表示为$\sigma_{yp}\sqrt{p}$ ，这里σ_{yp} 是调查变量Y 与应答倾向P 之间的方差，$\bar{p}$ 是调查样本的期望倾向值（Bethlehem，2002）。分因模型的方差为0，共因模型的方差为非0（不过在控制Z 的条件下为0），调查变量因模型的方差也为非0。

上述表述提醒我们，在一项调查中，无应答误差是调查参与者与要估计的变量之间是否有关的函数，随不同的估计值而不同。与之关

联的科学问题是，到底“哪些因素导致了Y与P的相关”或“是什么导致了调查变量与应答之间的关联”？

6.5 无应答现象解析

调查方法的研究发现，样本无应答可以分为三种类型，且由不同的原因造成。对于许多调查来说，会对调查统计量的质量带来不同的影响。这三类无应答分别是：

1）未能发出访问请求（“无法联系”，即未能与样本取得联系，例如在邮寄调查中遭到“地址不详”的退信）。

2）受访者拒访（例如，虽与样本取得了接触，但样本拒访）。

3）受访者没有能力提供所需要的数据（例如，受访者不懂问卷使用的语言）。

6.5.1 因递送调查请求失败导致的样本无应答

若采用某些方式进行调查，某些受访者的生活和行为方式决定了访员无法找到他们，这些样本被“弄丢”了。因此，由于无接触或未能发出访问请求，自然也就无应答。这里关键的问题在于受访者的“易联系性”，即是否易于让访员访到。图6.7试图列出影响受访者“易联系性”的各种因素。

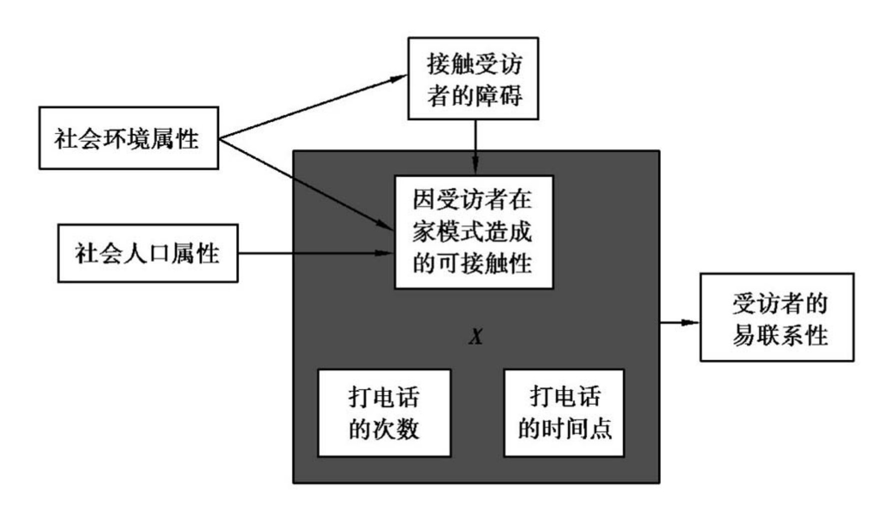


图6.7　影响样本家户“易联系性”的因素

在以家户为对象的调查中，如果我们知道家中什么时候有人在，那么访问一次就可以成功地与受访家户取得联系。事实上，我们不清楚什么时间能方便地找到受访者。因此，总是会要求访员对同一个样本家户多打几次电话尝试与之取得联系。有些样本家户安装了“防打扰”设施以阻止陌生人的到访（例如，公寓大门上锁以防止外人进入，使用电话答录机等）。有的人习惯将不熟悉地址的来信直接丢弃不看，如果采用邮寄问卷调查，这些受访者往往总是联系不上。有的人很少在家，这样的受访者就算访员多次给他家里打电话联系，往往还是联系不上。还有的家户电话使用了电信服务机构提供的电话免打扰服务项目，这样的受访者根本就没有机会知道访员曾多次尝试与他们取得联系。

例如，在全国药物使用与健康调查（NSDUH）中，2%的样本家户在筛选阶段未能联系上；在消费者调查（SOC）中，这种情况更为严重。在NSDUH调查中，如果受访者居住在单元公寓或其他有“防止打扰”设施的居住场所，则接触不到的情形就会没有规律。由于SOC采用的是电话调查，其联系不上的受访者特点有所不同，更多地集中于使用来电号码显示或其他拨入电话屏蔽功能的受访家户。

在实践中，如果前面的致电失败，越往后，取得成功的比例越小。举例来说，图6.8显示的是，为了成功获得首次接触所拨电话次数的家户数分布，数据来源于5项不同的家户调查。有的是电话调查，有的是面访，其中有些样本家户从头至尾，一直没能联系上。大约有一半联系上的家户是第一次致电就成功的。图中各条曲线的差异可能是抽样方案的不同以及访员在首次致电之后的致电规则差异所致。

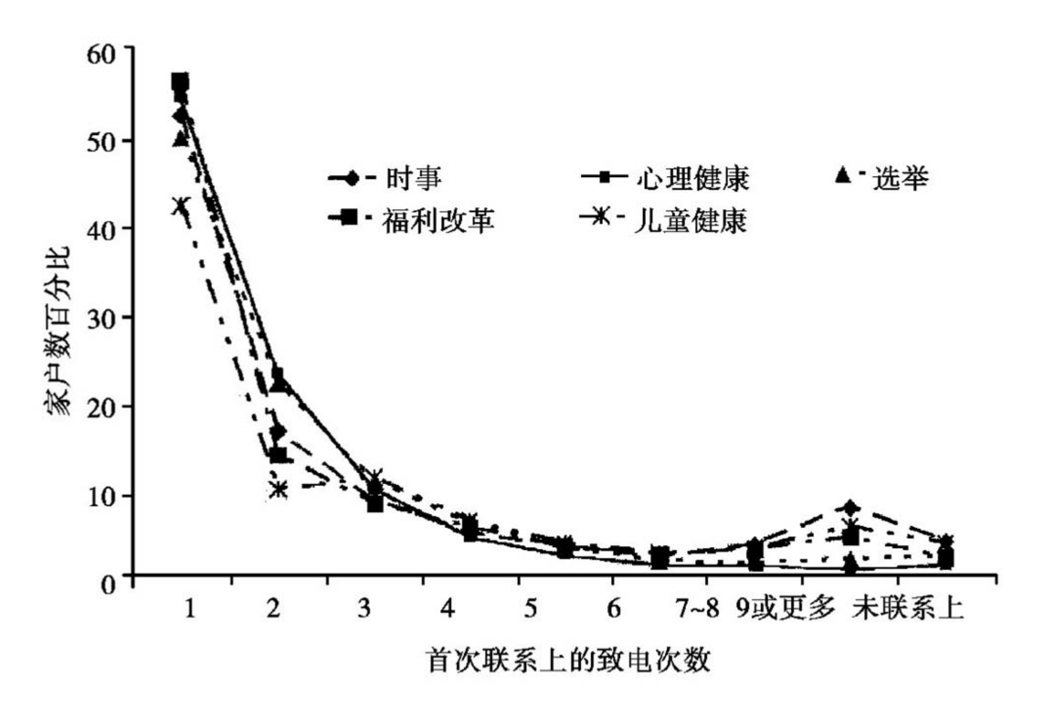


图6.8　五项调查中，为与受访家户成功取得首次接触所打电话次数的分布

（数据来源：Groves，Wissoker，Greene，McNeeley，and Montemarano，2001.）

怎样预知在家户调查中，到底需要打多少通电话才能与样本家户联系上？关键在于以下两点：

1）在晚上和周末致电要比在其他时间致电的成功率高。

2）针对有些人群，运用与其他人群不同的调查模式，可能更容易联系上他们。

被抽中的样本在家时，更容易让访员访到。那么，受访者什么时候在家呢？美国大多数家庭都有相当确定的起居规律。对于大多数在外工作的人，他们都会在固定的时间离家外出，每周如此。大多数在外工作的美国人，周一至周五上午8:00至下午6:00是不在家的。如果访员在这期间打电话到受访家庭，家中有人的可能性比较低。正如图6.9表明的，无论拨多少次，当地时间周日至周四的晚上6—9点，似乎都是致电的最佳时间段。这几个晚上的共同点是，第二天都是工作日。周五和周六晚上则不然，总地来说，在这两晚致电的成功率较低。在周末的白天致电要比在工作日的白天致电的效果好。根据我们的总结，在美国，晚上很少有家户空无一人的情形。

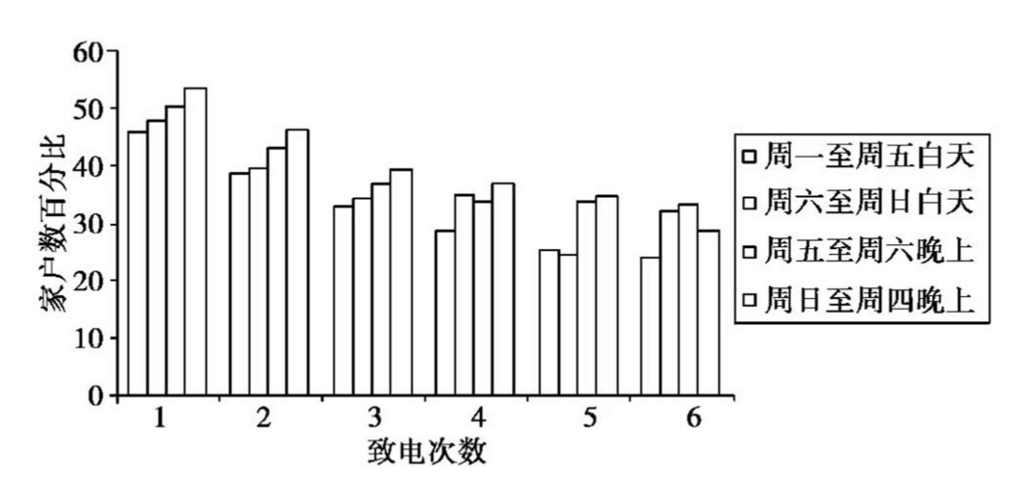


图6.9　原未取得联系的家户最终取得了联系的家户数分布，按不同致电次数和致电时间划分（数据来源：National Survey of Family Growth, Cycle 6）

在美国，访问成功率在各类人群间存在系统差异。联系成功率最高的是家中总是有人在的家户，其中又有多种情形，如家里有成员不

需要外出工作或有成员已退休，或家里有需要人照顾的学龄前儿童等情况。

与受访家户成功取得首次接触需打电话的次数，是测量与这类受访对象接触难易程度的度量指标。在家户调查中，要与设置了“防打扰”设施的家户取得首次接触，需要做更多次的努力。这些“防打扰”设施包括给公寓楼的主要通道上锁、给社区安装大门，或给乡下的住所安上有锁的大门等。在电访中，这些“防打扰”设施包括给电话设置来电号码显示，或配备其他具有呼入号码屏蔽功能的装置，这些都会造成与受访者取得首次接触需打更多次的电话。与乡下相比，与居住在城市的受访者取得首次接触也需更多次的努力（部分是由于许多城市受访者独居，且住处设置了进入障碍）。

邮件调查、电子邮件调查和用互联网的调查（都是不需要访员接触受访者的调查），一旦问卷发出，就意味着向受访者发出了他们随时可能收到的调查请求。也就是说，在邮件调查中，一旦一个家户收到了问卷，问卷就将一直待在那儿，直到家户成员对其作出处理。处理行为在一周的任何一天、一天的任何时刻，都有可能。与电访比较，这几种方式的调查与由访员试图与受访者取得联系的调查是不同的。第5章已经说明，这类由受访者自访的调查所具有的共同特点是：调查研究人员不易将受访者没看问卷情形与看到了但拒填的倾向进行区分。

简单地说，在访员指导下的调查中，各类人群不同的在家时间模式以及各家户不同的成员数量，会产生一些无应答，接触受访者的各类障碍会产生另一些无应答。各类人群不同的在家时间模式以及各家户不同的成员数量，在所有调查方式中都有影响；接触受访者的各类障碍，只对面访构成影响。例如，大门紧锁的社区或公寓楼，会对面

访造成困难，却不会对电访构成困难。故，对不同的调查方式，产生无应答的原因也各不相同。这就提示我们，在调查中，如果能将多种调查方式结合起来，将会降低无应答率。

对因未联系到受访者而产生的无应答，仍有一些有待解决的议题。因未联系到受访者产生的无应答发生与否，似乎与调查主题无关。这就是说，不是调查主题造成受访者难以联系，而是一系列其他因素造成受访者难以联系。无论调查的主题是什么，这些因素都同样发生作用。这就意味着，只有与这些因素相关的统计量才会有无应答误差。针对受访者难以联系的原因进行分析，可以为研究者带来指导，即在什么条件下未联系到受访者，进而带来无应答误差；在什么条件下，不会出现未联系到的情形。

因无联系无应答带来的估计偏差　当无应答与调查的统计测量联系在一起时，便会因无应答而带来偏差。举例来说，假设一项调查的关键统计量是独立生活的人拥有岗位性工作的百分比，如图6.10所示，其统计值便会受到无联系无应答的严重影响。从图中可以看出4项调查的无应答偏差。想象一下我们对单人户的百分比有兴趣，对这类统计量，用普查局的数据进行估计会得到很高的精度。如果采用电话家户调查，就会不成比例地漏掉那些不常在家的单人户。假设研究设计要求只打一遍电话，则低估的比例会高达27%～35%。如果打两遍电话，又会高估20%～32%。这个统计量对无联系误差非常敏感，因为无联系通常与样本对象在家的时长，进而能被访员捕捉到的机会有关。与之形成对照，对人们政治兴趣的估计就不大会受到无联系率如此大的影响。

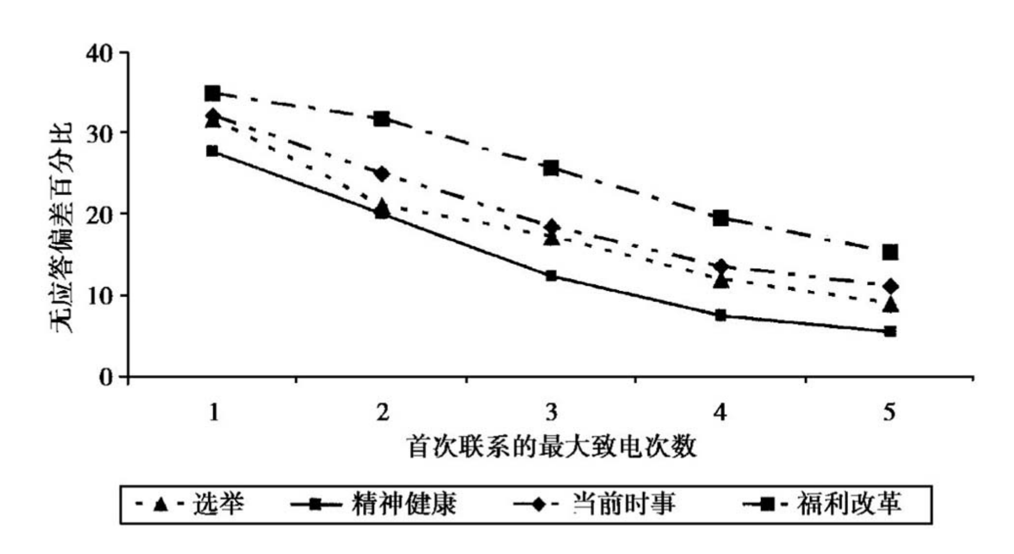


图6.10　估计单人户比例时无应答偏差百分比，4项调查要求的致电次数（数据来源：Groves，Wissoker，Greene，McNeeley，and Montemarano，2001.）

6.5.2　因受访者拒绝导致的样本无应答

已经讨论得比较清楚了的是，大量现实的困难决定了调查总是难以获得被调查者的合作（Groves and Kahn，1979，p227）。要获得成功的调查合作，需要受访者愿意对通过打电话、邮寄或登门拜访的完全陌生的访员所提出的调查请求给予积极响应。受访者还必须不惧怕陌生访客对其可能造成的生命和财产的伤害；不惧怕由于接受访问可能带来的名誉或心理上的伤害；还要对访员的保密承诺给予充分信任，并且相信这些访员能忠实反映他们的所思所想；对他们哪怕是最细微的想法，也能如实报告且不致招惹官司。对调查而言，请读者想想，哪里都会有这样和那样的困难！

社会调查的确普遍，对在大多数国家和地区的人而言，虽然其被访的频率还没有高到每天都有（或更频繁），但社会调查确实是越来

越频繁了。在某年曾有这样一项调查，即询问在过去的几年中是否曾作为受访者参与过除本次调查之外的其他社会调查。在1980年进行的这项调查中，有20%的受访者报告他们曾在前一年参加过，2001年则达到了60%。（这个数据有被高估的可能，因为其没有考虑到无应答者在过去几年的受访的情况。Sheppard，2001）。在美国，最常见的有以下两类调查：一类是由服务机构开展的顾客满意度调查；另一类，每两年的议会选举也会伴随着大量的社会调查，在选情激烈的选区更是如此。对大多数人而言，这两种情况之外的调查就比较少见了。

Groves和Couper（1998）认为，当人们被访员请求参与一项社会调查时，他迅速作出的第一反应是，想确认访员找他的目的是什么。因为与访员的相遇，被访家庭不是主动的一方，他们想知道访员到底想要什么。来访的陌生人为什么向我提出请求？请求我干什么？来访者来自哪个机构？对于这些问题，大家往往基于以前的生活经验已形成了一些惯常的答案，也相应地形成了应付这类请求的一些惯常的做法。相对于其他类型，社会调查的访问请求比较少见，所以受访家庭很容易将其与其他种类的请求混为一谈，譬如把社会调查类的电话当成了推销电话。在这种情况下，受访者可能因与社会调查本身无关的原因而拒绝接受访问。因此，社会调查经常需要再打电话过去，这是有效澄清自己不是推销电话的办法。如果某项调查是由一个著名的众所周知与营销事务无关的机构进行的，访员可以通过向被调查者强调这一点，以此把自己和推销员区分开。以下是社会调查方法的一些研究发现，这些研究发现证实了“受访者可能错误地理解访员的意图”。

“我不是在做推销”现象

有证据表明：受访者可能对访员的调查意图产生误会，因此在社会调查中，训练有素的访员往往这么做自我介绍——“您好！我是RCDF研究中心的玛丽·史密斯，这不是推销电话。我正在进行一项关于近期电话服务质量的调查。”很明显，这么说是为了防止受访者对访员打电话的意图产生误解。

这么说有用吗？对这一调查技术的评价好坏参半（van Leeuwen and de Leeuw，1999）。根据前文所述，受访者能否被打动，在于访员在自我介绍中传达的信息能否为对访员的信任感加分，而不是简单地来自“我不是在做推销”的申明。

1）受访者拒绝一项调查请求的决策，是在很短的时间内作出的（在电话访问中，大多数被受访者拒绝参与的决定是在前30秒内作出的）。

2）受访者第一次拒绝了调查，但当你再次与其联系时，他们经常就会同意接受访问（即所谓的转变率，常常为25%～40%）。

3）应当训练访员避免这种归错类的情况发生（如，在电话中尽早声明“我不是在做推销。”）。

受访者在很短时间内决定是否接受访问的事实表明：受访者的注意力集中在少数他们最关注的特征上。电访和面访的访员自我介绍，总是非常简短的（参见文本框 ）。有些受访者最关注的可能是调查的资助机构，另一些受访者最关注的是调查主题。在一些自访调查（如

邮寄问卷）中，受访者可以看见全部访题，此时，他们会特别关注调查的意图。

访员应该如何做自我介绍

要培训访员学会在快速简短的自我介绍中传达大量的信息。

例如，在电话调查中，“您好！我叫玛丽·史密斯，我在密歇根大学安娜堡校区跟您打电话，我们大学正在做一个全国性的研究项目。首先，我要确认我是否拨对了电话，请问您的电话号码是不是301-555-2222？”

或

在面访中见到受访者，“您好，我叫玛丽·史密斯，在密歇根大学调查研究中心工作，这是我的身份证件。密歇根大学正在做一项全国性的社会调查，我们想听听大家对一些事物的看法，如他们对经济的看法、对即将到来的总统选举的看法等。您应该已经收到了密歇根大学发给您的介绍信。”

为什么会出现样本无应答　样本无应答产生的原因，是调查方法的研究者们越来越关注的一个问题。学者们建立了多个理论框架来研究受访者参与问题，涉及4个不同层面的因素：

1）社会环境层面（如在大城市进行的入户调查中，往往遇到更多的拒访，有一个以上家户成员的家户相对于单成员家户，更

倾向于选择配合调查［Groves and Couper，1998］）。

2）受访者个体（如男性相对于女性更倾向于拒访［Smith，1983］）。

3）访员层面（如经验丰富的访员往往会比经验缺乏的访员能争取到更多受访者的合作［Groves and Couper，1998］）。

4）调查设计层面（如给受访者以物质激励会争取到更多的合作）。

前两个影响因素是研究人员无法控制的。例如有一些与调查完全无关的事项影响着受访者对调查请求的反应（如Tuskegee实验，对非裔美国男性做的梅毒调查，就经常用来作为低应答率的例子。参见第11.5.2节）。后两个影响因素，即访员和调查设计因素，是研究人员可以控制以提高应答率的（参见第6.6节）。

针对调查的参与性研究，人们使用了不少的理论，包括“机会成本”假说。这一理论认为，忙碌者之所以更倾向拒访，是因为他们花时间接受访问付出的代价比别人要高。有人还使用了“社会隔离”概念，社会隔离造成处在社会经济谱两端的人容易拒绝社会主流机构的调查请求。还有“主题兴趣”概念，即概念假定不同人群关注不同主题的调查，不同人群看待同一项调查的态度就有不同，这样就会带来关键估计值的无应答误差。另一个相关的概念是“过度调查”，指人群对过多的社会调查感到厌倦。这些概念在不同的学科里还有许多变体。糟糕的是，用其中任何一个概念来解释调查无应答，都不能得到令人满意的结果。大多数理论所提供的解释只涉及了“个体”或“社会环境”因素。

大多数理论都没有提及参与调查所带来的多元影响，如在决定参与那一刻的自我彰显。其中一个理论试图描述这些行为的心理基础，即“关注点杠杆理论”（Groves，Singer，and Corning，2000）。这个理论认为，不同的人在看待同一项调查时，关注点是不同的（例如调查主题、完成调查可能花费的时间、调查发起者、数据用途等）。对同一个关注点，有人会给出正面评价，有人却会给出负面评价。当然，对于调查人员来说，人们关注什么，是未知的。在争取受访者参与调查时，在与受访者的互动中，访员或问卷等到底为受访者提供了哪些信息进而让受访者作出是否参与的决定，是不知道的。受访者是合作还是拒绝，取决于他们关注哪些方面，以及对于这些方面他们给出了正面还是负面的评价。

图6.11中的两座杠杆代表两种不同类型的受访者，杠杆向右倾斜意味着受访者接受调查请求，向左倾斜意味着受访者拒绝调查请求。在收到调查请求之前，第一位受访者对调查主题给予最正面的评价，第二位受访者则对调查主题十分不感兴趣。第一位受访者的空闲时间很少，故对调查要花费的时间长短非常敏感；第二位受访者则不然。对调查支付的报酬，第一位受访者的看重程度只是中等，第二位受访者则非常在意得到现金报酬。如果访员在与他们进行接触时，重点强调了调查的资助方是谁（强调的程度以图中杠杆上各“砝码”的大小来表示）。则调查请求的结果是，第一位比第二位受访者更有可能接受调查。使用杠杆作为类比，受访者对调查特征关注程度的赋值，就像是杠杆上的砝码。在描述一项调查请求时，受访者对调查特征的评价就被称作“关注度”。这个理论带给我们以下几点启示：

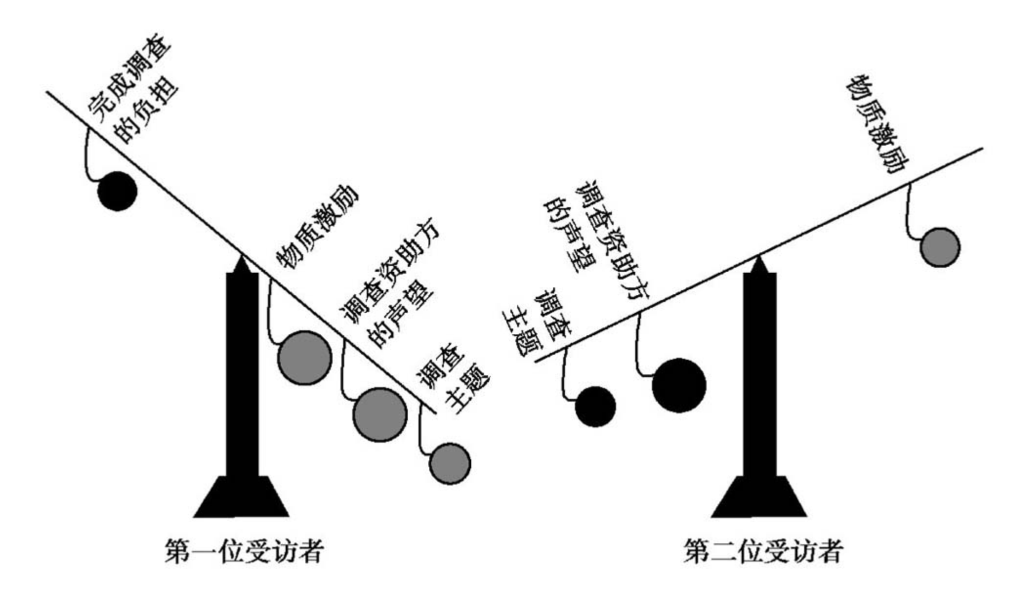


图6.11　对调查请求有着不同“关注点杠杆”的两类被调查对象

1）在提出调查请求时，访员并不清楚受访者会因什么而接受或拒绝。

2）没有一种调查请求方式适用于所有受访者。

3）访员必须学会了解受访者更可能对调查的哪些方面给予正面评价，并学会突出强调这些方面。

我们将在6.6节对此进行更详细的论述。

在无访员指导的社会调查中，调查设计要综合考虑采用各种文字和符号来关注调查属性的各方面。如在信封和纸张显见位置上显示调查资助方的名称，在封面信的显要位置附上支付的现金激励，还有，如把敏感访题放到问卷最后，这样可以使受访者在考虑是否接受调查时不会那么关注。

由拒访无应答带来的估计值误差　由拒访带来的误差，来自同样的逻辑。如果拒访与调查的关键统计量有关，那么，对估计值而

言，拒访率预示了无应答带来的误差。涉及这个观点的证据尽管不是十分充分，有一个比较研究的例子至少说明是否对受访者进行激励，是有差别的。医生协助的自杀是一个有争议的议题，在对该议题进行调查时，激励会提高对该议题没有兴趣的受访者的合作。其中的一个估计值是受访者是否参与争议（例如作为民间组织的成员或某相关政治事件的参与者）。在调查给予激励时，70%的受访者报告自己参与过。在调查不给予激励时，80%的受访者报告自己参与过（Groves，Singer，and Corning，2000）。如此，在调查不给予激励时，参与程度被高估了。

针对社会调查的拒访趋势已经变得越来越普遍。许多人都把注意力放在了图6.11的左边，因为他们发现这些因素与调查请求有关。例如，调查资助方声望的效应普遍存在，联邦政府的调查所得到了应答率就高于学术调查，学术调查又高于商业调查（Groves and Couper，1998）。完成调查的负担，在自访问卷中可以用问卷的页数来测量，页数越多，应答率越低（Goyder，1985；Heberlein and Baumgartner，1978）；在电访和面访中用时长来测量，其所显现的效应似乎不大清晰（Bogen，1996）。男性的拒访率常会高于女性（Smith，1983）。在所有现象中，城区是一个很好的应答率指标，越是城区，应答率越低（de Leeuw and de Heer，2002）。独自居住的成年人常会是拒访者（Groves and Couper，1998）；家有小儿的，应答率常会较高（Lievesley，1988）。当调查的关键变量与这些属性有关时，我们就能预测会产生基于无应答的无应答误差了。

6.5.3　因受访者无法提供信息导致的样本无应答

有时与受访者成功地取得了联系，他们也愿意接受访问，但却无法提供所需的信息。这种情况有多种原因：有时因为无法懂得我们使用的任何一种语言；有时因为他们不能理解我们提出的问题或无法回忆起相关的信息；有时是因为健康欠佳使得他们不能参与访问；有时是由于读写能力的限制，使得受访者不能参与像邮寄问卷这样需要他们阅读的调查。在商业类的调查中，受访机构也可能缺乏调查所要的特定格式或特定时期内的信息。

由于受访者无法提供调查所需信息的原因多种多样，故，由此产生的无应答对估计值的影响也各不相同。例如，在测量健康状况的社会调查中，因健康欠佳产生的无应答就会造成无应答误差。由于健康不良者的无应答是系统性的，因而对总体健康状况估计值就会偏高，但却不会对总体政治态度的调查估计值造成严重影响。

对这类样本无应答的发生原因和可能产生的影响，在方法论上的研究相对较少。在许多家户调查中，样本无应答较为少见。然而，在对老年人或移民的调查中，却常常发生样本无应答。对这类调查设计，研究发现，访员的作用、数据收集的方法、调查使用的语言以及样本特征都有影响。

6.6　减少样本无应答的设计特征

到这里，我们已经知道无应答误差会影响调查估计值的质量，影响的程度取决于导致无应答的原因与估计值的相关程度如何。在其他因素不变的条件下，无应答率越高，无应答误差出现的风险越大。总之，最直观的反应，无应答对调查质量造成影响的指标，就是无应答率，因此，调查方法论的学者们积极致力于降低无应答率的研究。

图6.12中将受访者参与调查的决策过程分解为三个步骤：受访者的易联系性、参与调查的初步决策和参与调查的最终决策。之所以会有“参与调查的最终决策”，是因为在许多调查方案中采取了多次说服措施等来争取不太情愿的受访者参与。

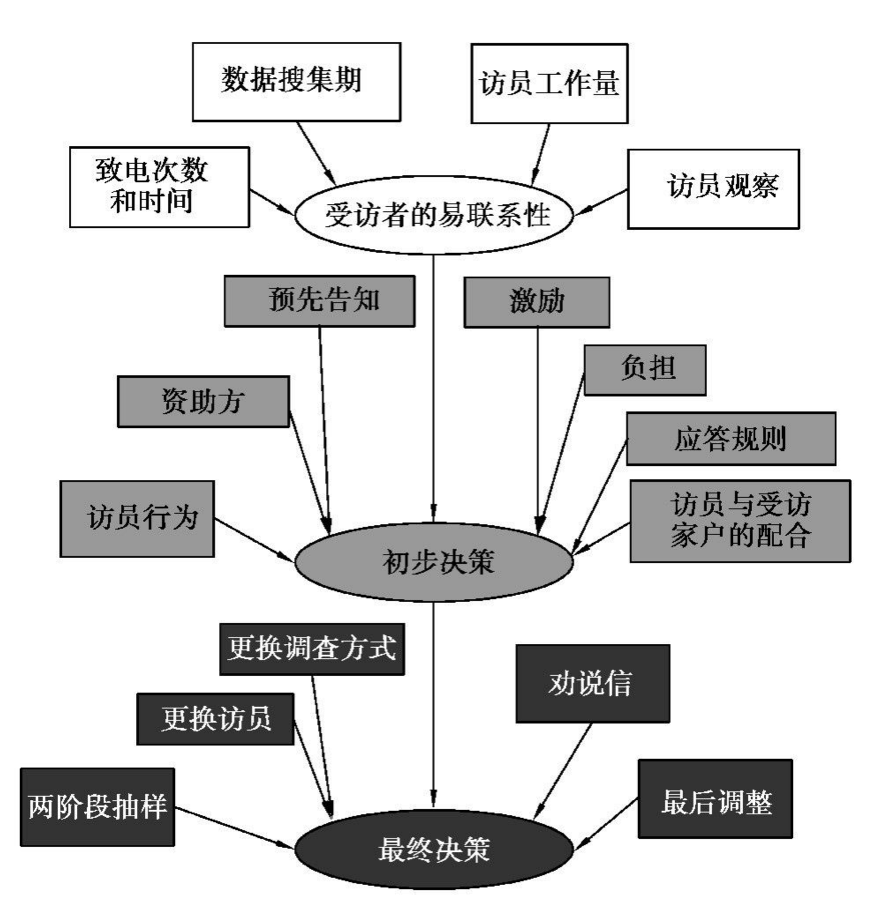


图6.12　降低样本无应答率的方法

在第5章，我们已经知道不同的数据搜集方式有不同的应答率。典型的发现是，在其他条件相同时，面访比电访的应答率高，电访比自访式邮寄和互联网访问的应答率高。在面访和电访中使用访员，会提高应答率，这不仅是因为使用访员会提高联系受访者的成功率，而且因为访员可以及时有效地获知受访者在考虑是否参与调查时的想法。

图6.12中有几处强调了访员的作用，从前面讨论的“关注点杠杆理论”可以得到关于访员行为的几个重要启示。这个理论告诉我们，不同的受访者面对调查请求时的关注点不同（即给同样的调查属性分配以不同的“砝码”）。访员尽量弄清受访者到底看重调查的哪些属性，就会有助于争取到他们的合作。

从“关注点杠杆理论”得到的进一步启示是，让访员给所有受访者同样的调查介绍，不会收到同样好的效果，经验研究已经证明了这一点。在Morton-Williams（1993）的文献中有一项试验，比较了让访员背诵同样的调查介绍和让访员自由发挥争取受访者两种情况，结果是，后者应答率更高（参见文本框关于Morton-Williams研究的介绍）。Grovers和Couper（1998）则认为，关于访员的行为，可以从两方面解释Morton-Williams的观察，即“保持互动”（maintaining interaction）和“定制”（tailoring）。有经验的访员善于发掘话题，进行互动，以吸引受访者注意力（无论这些话题是否与调查内容直接相关），“保持互动”可以让访员有机会了解受访者到底看重调查的什么，优秀的访员再据此“定制”与此有关的讲述。善于进行“定制式讲述”是有经验的访员，也是其比新访员赢得更多受访者合作的秘诀所在。有经验的访员会细心观察受访者的语言和行为来了解他们关注调查的哪些方面，再据此调整自身行为，他们对调查的介绍是按受访者的需要来“定制”的。

访员介绍如何接触样本家户

在焦点小组中，访员描述其怎样在准备访问样本家户。

“根据受访者的年龄、第一印象及其邻居等情况，准备不同的说辞。”

“在访问家户时，我会运用不同的技巧。运用色彩，有趣的别针、饰品，不是那些看起来很严肃、专业的东西，而是有趣的东西，如陶制的饰品，随便挂着的领带或围巾等。如果我访问的家户或其邻居家能从窗户看到有猫或狗，我会戴上有狗或猫的别针，等等。”

图6.12还表明，如果受访者最初不同意接受调查，通常要做进一步的努力，争取受访者的配合。这些措施包括更换访员、更换调查方式或给拒访者写信劝说。如果受访者最终仍拒绝合作（或所有争取受访者合作的努力均告无效时），则需要采取一系列其他措施。首先，可采用两阶段调查方案，换另一种调查方式跟踪那些拒访者。其次，还可以在数据分析阶段用统计方法进行数据的相应修正（这将在第10章讨论）。

本节剩下部分将就有关如何提高应答率文献的结论进行简要的概括。相关的文献量很大，包括一些研究著作（Goyder，1987；Brehm，1993；Groves and Couper，1998；and Groves，Dillman，Eltinge，and Little，2002），以及上百篇学术期刊论文。这些研究多采用随机试验对两种不同调查方案的效果进行比较。对一部分受访者采用一种方案调查，对另外一部分受访者采用另一种方案调查。应答率高的那种调查方案就被认为是更好的方案。有时研究人员可以获取受访者特征的其他信息，这样就可以评判提高应答率是否有助于提高调查估计值的质量。然而，这些研究的目的，通常在于发现如何提高应答

率，而不管这些措施是否有助于提高估计值的质量。我们按图6.12的框架来进行以下的讨论。

尝试与受访者取得接触的次数和时间 多项研究表明，在受访者自访和访员指导两种调查方式中，尝试与受访者取得接触的次数和时间段越多，与他们成功接触的可能性越大。Goyder（1985），Heberlein和Baumgarther（1987）的文献研究结果表明，争取与受访者取得联系的持久努力会有助于降低无应答率。在美国开展的电话调查和面访调查表明，周日至周四的晚上（即工作日前一晚上）和周末的白天较容易和受访者取得联系。只有一小部分受访家庭仅在周末的白天才能联系上。

Berlin, Mohadjer, Waksberg, Kolstad, Kirsch, Rock, and Yamanoto（1992）

关于激励受访者和访员调查成功率的论述

Berlin等人发现，给受访者激励能降低调查总成本。

研究设计：对投票地点采取概率抽样，对投票者采用系统抽样。在“全国成年人读写能力调查”读写能力测试的试调查中，对样本随机给予如下3种激励：不给激励、承诺给20美元、承诺给35美元。访员先询问问卷中有关受访者背景的访题，然后由受访者自己完成读写能力评估问卷。抽样方法是，对人口普查300个区划的大约2 800个家户实行区域概率抽样，在每个区划中随机采用某种激励。

研究发现：激励把受教育程度较低的、以前大多不愿接受这类读写能力调查的人群“拉进”了应答者队伍。如下表所示，激励为35美元的赢得了最高应答率，没有激励的，应答率最低。

	激励		
	无激励	20美元	35美元
应答率	64%	71%	74%
访员成本	\$130	\$99	\$94
激励成本	0	20	35
总成本	\$130	\$119	\$129

由于节省了访员在争取每个受访者上花费的时间，20美元的激励与不采取激励相比，的确物有所值。35美元的激励虽然获得了最高的应答率，其成本收益率却不是最好的。

研究局限：由于这项调查主题不大常见（即读写能力的自访问卷），并由政府资助实施且调查区域范围很小（16个主区），进而限制了其结果的应用。受这些因素的影响，标准差或很不稳定。

研究意义：给受访者以激励反而会节省调查总费用的研究发现与人们通常的直觉相悖，具有重要意义。

数据搜集期的长短　数据搜集期越长，受访者就越有可能收到调查的请求。搜集期多短算太短？这里值得一提的是目前人口调查的数据搜集期仅有10天，却获得了高达100%的样本接触率。这个例子表

明，借助在调查中配备高素质的访员队伍，在相对较短的时间内也可以与大部分受访者成功接触。邮寄问卷调查的数据搜集期较长一些，这是由于邮寄问卷本身需要一些时间。

访员工作量的大小　布置给访员的每个个案，都需要访员耗费一定的时间与受访者接触。在家户电访中，使用通常的拨号规则一般只有50%左右的首次接触成功率，若给每位访员布置的任务量太重，访员花在每个家户上的平均时间就会不足。Botman和Thornberry（1992）指出，若调查时间太紧或样本量太大，无应答率、无接触率（由于和受访者取得接触的努力不足）和拒访率（由于争取拒访者回心转意的努力不足）会升高。

访员观察　在与样本进行接触时，面访有一个明显的优点，即访员能够直接观察受访者的特征。有时，会观察到一些明显的信息点（如院子里的儿童玩具暗示家有儿童）；有时，受访家户的邻居也会提供受访者的有关信息。访员间接搜集受访者的相关行为信息有助于有效掌控调查进程。如果受访者问及与调查本身相关的问题，往往预示着他们最终会接受调查（Groves and Couper，1998）。

资助方　在世界上的多数国家，政府组织的调查往往比学术性或私人部门的调查获得更高的应答率。当调查资助方与受访者有某种联系时（如组织成员），这种联系的密切程度与应答率是相关的。例如，若将待访样本随机分配给美国人口普查局和密歇根大学调查研究中心进行调查，分派给人口普查局样本的拒访率是6%，而分派给密歇根大学调查研究中心样本的拒绝率是13%（National Research Council，1979）。许多学者认为，政府组织调查的应答率之所以高，是由于大家普遍认为政府机构需要居民的信息是合情合理的，而且有益于受访者。

预先通知　　预先发一封信通知样本家户，比不预先通知能够带来更高的合作率（参见Traugott，Groves，and Lepkowski，1987；相反的结论参见Singer，Van Hoewyk，and Maher，2000）。然而，通知信的效果取决于其署名机构，也许使用什么样的文具也会影响应答率。例如，在一项随机实验中，相同内容的预先通知信，用市场调查机构信笺所获得的应答率就比用大学研究机构信笺的应答率低。实际上，由市场调查公司署名的信比不写信收到的应答率都低（Brunner and Carroll，1969）。访员倾向于认为，预先通知很有意义，因此，在学术调查中预先发一封信已经成为惯例。

激励受访者　　给受访者提供一份附带的好处，可以提高应答率（Singer，2002）。现金激励比相同价值的其他激励效果更好。在向受访者提出调查要求之前就提供激励的效果比完成调查后才提供激励效果好。激励程度越高，效果越好。也有研究表明，激励程度越往上，激励效果存在着递减的趋势。在物质激励效果足够高的情况下，由于此时支付给访员的报酬或追踪调查的成本相对变低，就会使调查的总成本下降（参见文本框 ）。

Morton-Williams（1993）
论访员定制

Morton-Williams（1993）提供了一项在英国进行随机实验的研究结果，肯定了访员根据受访者的关注点，对调查介绍进行应变性调整的意义。

研究设计：要求30位访员中的14位在面访的介绍中照稿宣读，其余的16位则根据自己对受访者的观察判断来决定自我介绍的措

辞。要求所有30位访员在介绍时必须说出自己的姓名、来自哪个组织、出示身份证、提及组织的独立性、对调查进行介绍并解释受访者的地址是如何被选上的。照读的介绍稿，则有着确定的、访员要用的每一个词语。

研究发现：照稿宣读的访员获得了59%的应答率，未照稿宣读的则有76%的应答率。

研究局限：关于研究过程的描述并未明确访员或受访家户是如何分配的。因此，可能会把受访者之间真正的差异与处理方式的差异相混淆。统计分析似乎并未解释因访员的不同而带来的那部分变异；由于实验中的访员人数很少，故不清楚应答率的差异是否超过了随机差异。

研究意义：这项研究第一次作为证据支持了访员在调查的开始阶段的介绍时，照稿宣读可能有损应答率的看法。

应答负担　有各种证据表明，完成调查所需时间和应答负担大小都会影响应答率。同样有证据表明如果受访者感到所需时间过长或需要他们自访的问卷很复杂，也会降低应答率。例如，在总结大量的调查情况之后，Heberlein（1978）和Goyder（1985）发现每增加一页自访问卷，应答率就会降低0.4个百分点。

应答规则　允许样本家户其他成员应答比只允许某个被抽中的具体成年成员应答，有更高的应答率。同样的，允许他人代答比只允许受访者本人应答有更高的应答率。如此，有一些误差指标，例如，在健康调查中，如果允许代答报告损伤和疾病的比例就会偏高。

访员介绍行为　尤其是在电访中，访员与受访者沟通最初几秒的效果好坏会影响应答率。关于这个议题，实证证据并不多，不过，已有的研究表明，沟通时的语调变化和语速与较高的应答率相关（Oksenberg，Coleman，and Cannell，1986）。Morton-Williams（1993）的研究显示，如果访员在自我介绍中使用一成不变的刻板措辞，则遭到受访者拒绝的可能性较大（参见文本框 ）。

访员与受访家户的匹配　这方面并没有相关的研究，据推测，将访员与受访家户进行匹配可以增进受访家户对访员的信任感，使访员易于被接受，从而提高应答率。例如，Nealon（1983）的研究提及在一项针对农场工人的调查中，女性访员比男性访员获得了更高的应答率。在人类学的调查实践中，普遍采用“土生土长”的访员也是同样的道理。在考虑无应答误差和这种实践时，在考虑因访员属性会改善受访者参与的同时，是否也会影响受访者的应答（第9章的部分内容，说明访员的性别和种族会影响到受访者针对性别和种族访题的应答）。

更换访员　如果访员最初遭到了某位受访者的拒绝，督导常见的处理办法是，更换另一位其某些特征更有希望为受访者接受的访员（如前所述，在某些特征上进行匹配）。新更换的访员预先了解被拒过程，包括相关的文件资料，然后再一次试着与受访者接触。这种做法，也可被看作非常自然的“随机应变”，即通过对访员特征随机应变的调整来适应受访者。例如，在面访中，男性访员可能会让一个年迈独居的女性受访者感到害怕，如此，就可以考虑更换为一个年龄较大的女性访员。

更换调查方式　许多研究设计都在初期阶段先采用花费较少的数据搜集方式（如邮寄问卷），然后对在第一种方式中没有应答的受

访者，使用花费较多的方式（如电访或面访）。如果调查只能使用一种方式，显然面访比电访或邮寄问卷的应答率都高。将几种调查方式合并使用，就可以实现资源的最优化，提高应答率。

劝说信　对于初次拒访的受访者，一般的做法是，发一封信，强调说明调查的重要意义，并说明访员会再次致电府上征询其想法。实践中，一般根据受访者的关注点对信件内容进行调整（如给担心个人隐私受侵犯的受访者在信中强调调查的保密性）。

从上面简短的归纳中我们可以看出，有许多方法可以降低无应答率。其中有些方法有坚实的研究结论为基础，有些方法则基于调查的经验总结。并不是所有的方法都适用于所有的环境。也有因拙劣地照搬他人的经验却失败了的例子。社会调查实践者面对的挑战就在于，如何在研究目的和目标人群给定的情况下，制订出最有希望成功的调查策略。

图6.12中最后两个方框指的是，以两种不同的统计技术为基础的调查工具。用*两阶段抽样*（two-phase sampling）方法获得无应答者的概率样本，采用新方法与之联系并邀请其参与。用从样本获得的数据来估计所有无应答者的特征。*调查后调整*（postsurvey adjustment）（见第9章），就是使用现有应答者的数据，对无应答者的缺失数据进行补救的统计技术。例如，如果城市地区的应答率较低，在数据分析时，相对于农村应答者而言，给予城市应答者数据较大的权重。

尽管关于如何提高调查应答率的文献成果丰富，这个领域仍然有许多尚未解答的问题，如：

1）如果我们成功地争取到了一位原先不愿意参加的受访者的合作，他们给出的应答是否会带有更多的测量误差？

2）在什么情况下提高应答率的努力会降低无应答误差，什么情况下不会？

3）怎样使我们的努力和经费在降低非接触率和拒访率之间平衡地分配？

4）在调查预算给定的情况下，兼顾抽样误差和无应答误差，什么情况下研究者不必只在降低无应答率上下功夫？

研究无应答，对社会调查领域的未来发展非常重要，因为在目前做研究设计时，会投入很大一部分预算用于减少无应答率，具有实践指导意义的科研理论却寥寥无几。

6.7　选项无应答

上面的讨论集中在样本无应答（即从某个样本未得到任何调查结果）。在某些类型的调查中，还存在着严重的选项无应答。选项无应答指的是，问卷的某个访题没有应答，如在消费者调查（SOC）中，受访者愿意接受调查且开始应答，可当访员问及其去年的家庭收入时，他们却会拒绝回答。

选项无应答与样本无应答一样，会对调查估计值造成影响，只是，选项无应答仅在使用无应答访题对应的数据进行统计计算时，对

相关估计值造成影响。因此，本书此处 的公式，实际上反映的是样本无应答和项目无应答给误差带来的综合影响。

各种迹象表明，影响选项无应答和样本无应答的因素不同。受访者决定不参加调查的决策主要基于对调查的描述，选项无应答则是在受访者完全了解调查的情况下出现的。从方法角度进行的研究表明，影响选项无应答的因素有以下几个方面：（a）受访者对访题的内容缺乏足够理解；（b）受访者不能回忆起与访题相关的信息；（c）受访者缺乏提供信息的意愿或动机（参见Beatty and Herrmann，2002；Krosnick，2002）。然而，对选项无应答的研究仍处于初期阶段。大部分涉及问卷访题措辞的研究（详见第7章 ）认为，访题会影响受访者的应答。针对选项无应答影响因素的研究，对调查实践具有极大的指导意义。

有证据表明，选项无应答源于受访者认为自己无法给出足够准确的答案。实验研究表明，人们往往不愿给出一个确切的收入数据，却愿意估计一个收入范围（如收入在50 000～75 000美元［Juster and Smith，1997］）。对受访者应答动机的研究表明，开放式访题（需要应答者自己提供访题的答案）要比封闭式访题（应答者从给出的应答选项中选择）更易于出现访题无应答。

图6.13是由Batty和Herrmann提出的受访者应答过程的认知状态模型，该模型将受访者对于调查访题总体的认知状态划分为4个层次：已有的、可获知的、可估计的和不可估计的。这4个状态是依受访者可提供与访题相关信息的能力来排序的。Batty和Herrmann认为，这些信息会有*随性误差* （errors of commission，即受访者在不掌握足够相关信息情况下，对访题作答）和*遗漏误差* （errors of omission，即受访者在掌握足够相关信息情况下未作应答）。如果受访者受到社会影

响，其应答也会给数据带来测量误差。出现选项无应答有时是正常现象（如在认知状态处于“不可估计”时），有时却是一种响应误差（如认知状态处于“已有的”时），后一种情况也可能出现在下面的情形中，如受访者希望受到社会赞许而拒绝回答某个访题（或说“我不知道”）而不给出有可能不受社会赞许的真实应答。

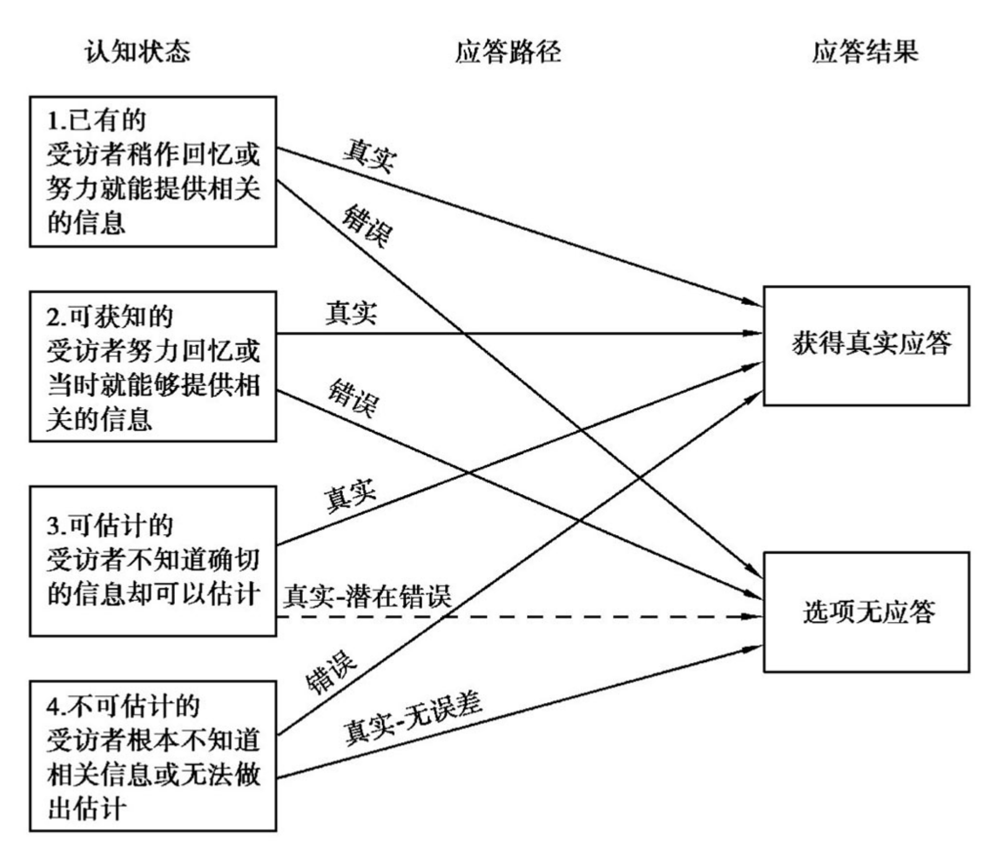


图6.13　Batty和Herrmann的受访者应答过程之选项无应答

减少选项无应答的工具就是减少任何单个访题的负担，减轻心理压力，提高对受访者隐私的保护（如让受访者自访），以及让访员对访题作进一步澄清或针对受访者的应答进行追问。解决选项无应答的策略通常与解决样本无应答的策略不同。针对选项无应答，正如前面的实例所示，研究人员通常可以其他相关信息来修正和调整。因此，

通常用补值法来解决选项无应答带来的数据缺失问题，采用调整群间权重的办法解决样本无应答的问题（该内容详见第10章 ）。

6.8　无应答的偏向性与其他误差来源有关吗？

在我们结束这一章之前，还需要告诉读者，一些调查方法研究表明，无应答或许与其他一些误差有关，尤其是覆盖性误差和测量性误差。从人口普查文献来看，我们发现年轻的单身男性常与多个家户有着弱联系（故容易在地址抽样中被遗漏），也容易成为无应答者。类似的，有证据表明，最不情愿应答的人，在应答时也最不连贯，进而有更大的测量误差（参见Olson，2006）。尽管这是最近出现的一些研究，却也进一步提醒我们，在防止无应答对调查估计值产生的影响时，不是简单地提高应答率就可以了。

6.9　小结

社会调查就是通过对少数样本进行测量和估计，来描述更大数量总体的特征。当不能对样本进行完全测量时，统计量的计算只能基于应答样本，从而使调查统计量的质量受到威胁。有两种类型的无应答，即样本无应答和选项无应答。

计算无应答率最简单的公式是，未测量的合格样本数与合格样本总数之比。在实践中，无应答率有时很难计算，原因是，常常会不知

道未调查到的样本是不是合格样本，而且在抽样设计中，不同抽样框要素往往有不同的抽样概率。

并非所有的无应答都有损于调查估计值的质量，其中与关键调查统计量计算相关的无应答是最有害的（如在调查人们如何利用自己的时间时未能与很少在家的人取得联系）。这类无应答被称作“不可忽略的”（nonignorable）的无应答。无论是描述性或分析性的统计量，无应答都可能对其质量带来损害。同一个调查中，不同估计值的无应答误差可能大小有所不同。

各种各样的原因造成了三类样本无应答，且以不同的方式影响调查统计量的质量，分别是：由于未能接触到受访者而引起的样本无应答，由于受访者拒访产生的样本无应答，以及由于样本没有能力提供相关信息导致的样本无应答。

无应答率自己不会预测其对单个调查估计值产生的无应答误差。相反，有证据表明，即使在同一项调查中，不同估计值的无应答误差之间差异很大。尽管人们长期忽视无应答有影响或无影响的条件，却始终在调查中嚷嚷着要提高应答率。大多数专业调查机构的指南，都描述了提高应答率的方法。

社会调查研究者们掌握了许多提高应答率的方法。包括：进行多次尝试、延长数据搜集期、减轻访员工作量、借助公众信赖的机构调查、让问卷简洁、允许他人代答、根据受访者的关注点调整访员行为、让访员和受访者特征匹配、初次被拒访后写信劝说、更换访员和调查方式来争取不太情愿的受访者参与调查，以及采用两阶段抽样法。上述所有方法几乎都要花更多的时间或精力去联系受访者或与他们沟通。总体上都会增加调查成本。

对调查研究者而言，还有一个重要的挑战，就是怎么知道无应答在什么情况下会影响调查统计量的质量、什么情况下不会？对此，需要有更进一步的研究。不进行这样的研究，我们就不能保证在调查中投入巨大努力来提高应答率是明智之举，也就不能肯定地说，应答率不高也无关大局。

关键词

样本无应答（unit nonresponse）

无应答偏差（nonresponse bias）

不可忽视无应答（nonignorable nonresponse）

定制（tailoring）

调查后调整（postsurvey adjustment）

访题无应答（item nonresponse）

应答倾向性（response propensity）

保持互动（maintaining interaction）

两阶段抽样（two-phase sampling）

进一步阅读资料

Groves, R., and Couper, M. (1998), *Nonresponse in Housegold Interview Surveys* , New York: Wiley.

Groves, R., Dillman, D., Eltinge, J., and Little，R. (eds.) (2002), *Survey Nonresponse* , New York: Wiley.

Särndal, C., and Lundström, S. (2005), *Estimation in Surveys with Nonresponse* , New York: Wiley.

作业

---

1. 以下是一项以电话访问为调查方式的最终样本。调查样本的生成方式是，在全美大陆48个州（不包括阿拉斯加和夏威夷）加哥伦比亚特区的全部电话号码中随机抽取2 127个电话号码。调查的主题是“家庭废弃物回收行为—— 家庭所在城镇废弃物回收装置的可得性和家庭对废物回收装置的利用”。所有样本号码最多拨打20次，拨打时间包括周一到周五和周末，白天和晚上，直到拨打成功。

完成样本	614
拒访样本	224
每次打过去都是电话答录机应答	180
每次打过去都没人接	302
和家户联系上了，因拒访外的其他原因未访	127
商用电话/无人居住	194

无效号码 486

请以下列三种方式计算应答率（可参考http://www.aapor.org上的“Survey Methods”部分）：

（a）假设所有未接电话号码都是合格样本

（b）假设所有未接电话号码都是不合格样本

（c）估算未接电话号码中合格样本的比例

2. 假设要从全国范围内随机抽取1 500所乡村学校的课程系统进行调查。调查旨在研究影响乡村学校是否设置性教育课程的因素，并通过调查，比较坐落在“宗教保守群体”（CRG）占居民大多数的地区的学校和坐落在“宗教保守群体”（CRG）占居民少数的地区中的学校之间，在课程设置上的差别。

学校系统总体	样本规模	应答率	应答学校系统	提供性教育课程的百分比	
				应答者	无应答者
主要是 CRG	500	50%	250	5%	0%
不主要是 CRG	1 000	60%	600	50%	35%

假设两类样本的样本规模比例代表了总体两类样本的真实比例（即等概率样本）。那么，要估计设置了性教育课程学校总体的比例数，根据相应的调查结果，计算估计是36.8%。计算过程如下：

CRG占居民大多数：5%×250=12.5

CRG占居民少数：50%×600=300

12.5+300=312.5

312.5/850=36.8%

请问以上估计设置了性教育课程学校比例数的计算中，无应答误差的大小。

3. 在本章中讨论过给受访者以激励能够提高应答率。

（a）一般情况下，在提出调查请求之前就付给受访者激励比先承诺激励等调查结束后再兑现的做法，能带来更高的应答率，请阐述这种现象背后的逻辑。

（b）请论述为什么给受访者激励往往会使调查总成本降低?

4. 你们即将完成一项针对某专业组织成员的电话访问，调查主题是，他们为支持组织的发展付出了多大努力。当前的调查应答率是80%，你们面临的问题是，即是否要投入一部分研究经费以进一步提高应答率？调查要测量的关键指标是，这些成员参加组织各地方分会月度例会的人数百分比。在80%应答率的情况下，指标估计值是，42%的会员参加月度例会。

（a）事实上，参加月度例会的会员比例有没有可能超过一半?

（b）对无应答者，其出席月度例会的情况会有怎样的特征?

5. 就总体均值而言，请简述其无应答率和无应答偏差之间的关系。

6. 你们正在计划进行一项健康照料的研究，研究总体是去年参加美国游泳大师赛的参赛者。调查访题涉及饮食和锻炼方法，同样的访题也曾在最近一项针对全美健康锻炼行为的调查中使用过。通过随机实验，你发现若对受访者采取每份问卷10美元的激励，应答

率会提高20%。请举出理由说明，在什么情况下，激励能收到这样的效果；再举出理由说明，在什么情况下，激励会收到与此不同的效果。

7. 假如你们已经掌握了调查前使用预先通知信会提高应答率的知识，请为以下3项社会调查分别选择写预先通知信的信纸类型。

（a）假设目标是使应答率最大化。请在两种信纸间作出选择，使用印有资助组织名称的信纸（该组织资助该项调查），还是印有数据收集组织名称的信纸（该组织负责该项调查的数据搜集），然后说出理由。

（b）假设目标是无应答误差最小化，又该如何作出选择？请说出理由。

资助方	数据搜集方	目标总体	用哪个组织的信纸写告知信?	选择的理由
商业信贷公司	学术调查中心	美国家庭		
联邦政府	商务营销研究公司	低收入家庭		
高速公路建设游说组织	非营利的研究组织	高速公路建设公司的总裁		
商业信贷公司	学术调查中心	美国家庭	信用卡欠款均值	
联邦政府	商务营销研究公司	低收入家庭	喜欢食品券的比例	
高速公路建设游说组织	非营利的研究组织	高速公路建设公司的总裁	认为游说有效率的比例	

8. 如果是一项对数据搜集期有严格约定的家户调查，在你的国家，采用电话访问方式，哪个子总体会是联系不上或无应答的家户？为什么？

9. 使用关注点杠杆理论描述这样的情境，即调查资助方会影响受访者对政治候选人支持率的统计量。

10. 在研究激励效应时，针对大学毕业生进行了一项实验。实验将样本随机分为两组，一组收到的访题涉及政府负担的处方药，另一组收到的访题涉及环境问题。每一个小组，又一分为二，一组没有激励，另一组有5美元的激励，由此形成了一个2×2的实验设计。

    简要描述这个实验可能的发现，即激励与主题之间的关系。

11. 列出3条理由，说明为什么一般而言面访比电话访问调查会有较高的应答率。

12. 本章讨论过在调查中提高应答率的备选方法。

    （a）列出提高应答率的3种方法。

    （b）在成人家户总体中，想一想上述3种方法中，哪一种方法对其中某个子总体有效而对其他的子总体无效，并作简要解释。

13. 本章的讨论说明，降低无应答率有时候会增加某种简单统计量如均值的无应答误差。设想，如果对一个组织（譬如舞蹈小组）进行调查，希望了解组织成员参与组织活动年头的平均值，请问哪种调查设计特征会产生类似的问题？

7　调查中的访题与应答

正如在本书第5章中指出的那样，调查使用了多种方法来搜集受访者的信息。

其中最常见的方法，就是问卷（questionnaire）调查，即用问卷作为一套标准访题，对受访者进行调查。问卷的访题，通常以某种格式排列，且有应答选项。一般情况下，问卷调查由访员依据问卷进行提问，受访者根据访员的提问作答。也有许多调查，会让受访者自己根据问卷访题来作答。在我们引用的6项调查中，其中有3项（NCVS，SOC以及BRFSS）几乎完全依赖访员的访问，另外3项（NSDUH，NAEP和CES）则既有访员调查的部分，也有受访者自访部分。在过去20年左右的时间里，调查问卷逐渐转化为电子形式，利用计算机向访员或受访者显示问卷的访题。不过，无论访员是否介入受访者的应答过程，无论是纸版问卷还是电子问卷，大多数调查依然极大地取决于受访者对调查问题的理解，以及给出的应答，这才是问卷调查搜集的信息。

7.1　调查测量的替代方法

调查并不总是依赖于受访者提供的应答。例如，许多调查从商业或其他机构收集信息，且常常运用机构的已有记录来获取信息。在这种情况下，问卷的作用，与其说是访问的脚本，不如说是数据记录的表格。同时，访员调查的对象是已有的记录，而不是受访者。CES有时

更接近于这种模式。同样，教育调查也可以用学生成绩数据来补充问卷调查数据；健康调查也可以从医疗记录中获得相关信息，而不是完全依靠受访者对诊疗活动的应答。不过，即使信息主要来自于已有的记录，受访者对信息的可及与获取也仍扮演着重要的角色。有些调查虽然是通过访问来搜集数据，却要求受访者提前搜集相关的记录以帮助其作出准确的应答。例如，国家医疗开支调查（National Medical Expenditure Survey）及相关的医疗开支追踪调查（Medical Expenditure Panel Survey），都鼓励受访者保留医疗账单和其他记录来帮助回答诊疗及费用相关的问题。有时候，这些记录对受访者的应答极其重要，但常遇到的情况是，在需要的时候，这些记录就是没有。例如，极少有家庭会保留日常支出的详尽记录；如果留着这些记录的话，那么在遇到美国劳工统计局的消费者支出调查（Consumer Expenditure Survey，CES）询问家户支出的时候，就很有用处。这一类的调查，都试图说服受访者保留日常生活相关事件的记录。对依赖既有记录的调查而言，日志调查把受访者的记忆负担转向了记录的保留。

调查中，另一类测量则与标准化的心理测试有关。许多教育调查都试图把教育产出与学生的父母、老师以及所在学校的特征进行关联；为了进行比较，这类研究通常会对学生进行认知测量。调查范例之一的全国教育进展评估（NAEP），就非常依赖于学术成就的标准化测试。

这一章，集中讨论由调查问卷引出的问题。理论上，所有调查都要使用问卷，即使有些调查不使用问卷，也要依靠诸如记录提取表格或日志之类的标准化数据搜集工具。编制和测试问卷的许多原则同样适用于其他的标准化工具。

7.2　应答访题的认知过程

几乎所有的问卷调查都需要受访者对访员的提问进行应答，或自己对问卷上的访题进行填答。一些研究者曾试图弄清楚由访题引发的的受访者心理活动，大多数研究结果都包括了以下4组步骤：“理解”（受访者解释访题）、“检索”（受访者回忆应答访题所需要的信息）、“判断”（受访者联系或归纳搜寻到的信息）以及“报告”（受访者形成应答并用规定的格式输出），图7.1展示了这个过程。

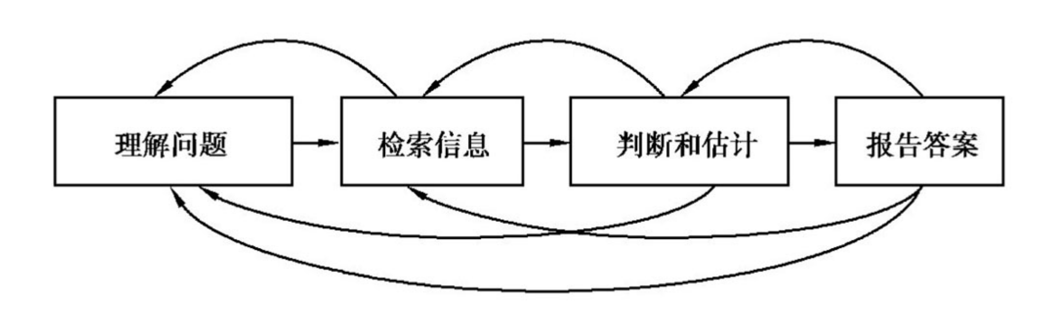


图7.1　调查应答过程的简易模型

一些情况下，在访问以前，把受访者首次经历访问的认知过程纳入考量非常重要。“编码 ”（encoding）是从经验形成记忆的过程。通常，调查设计者对这些过程几乎没有影响，却有充足的证据表明，如果调查设计者考虑到受访者如何编码访题追寻的信息，就可以极大地改进访题。

最后，在自访问卷时，受访者也会形成自己的填答方式，确定哪是下一个应答，并理解填答提示语。理解填答线索和提示语的过程，就是“理解”步骤的一部分。（调查反应过程模型的早期例子，参见Cannell，Miller，and Oksenberg，1981；以及Tourangeau，1984。

较近期的使用，参见Sudman，Bradburn，and Schwarz，1996；以及Tourangeau，Rips，and Rasinski，2000。）

对应答时复杂和严谨的认知过程，人们很容易产生误解。Krosnick（1999）以及Tourangeau，Rips和Rasinski（2000）还有其他研究者对应答过程的详细研究表明，受访者常会略过图7.1的一些步骤，粗心地完成其他步骤。访题常常难以应答，需要受访者使劲回想并作出判断。不妨看看来自调查的实例，照搬来的原问卷访题。比如，NCVS问卷的空白部分需要填上具体日期。

> 平均来说，在过去的6个月中，即，从________开始，您购物的频率？比如，去药店、服装店、杂货店、五金店或便利店。［NCVS］

> 如果把整个国家的商业环境当作整体，您认为在接下来的12个月，我们会遇到好的经济？不好的经济？还是其他的什么？［SOC］

想象一下，对NCVS提出的购物访题，要给出确切的应答有多难。幸运的是，NCVS寻求的是大致的频率，如“一周一次”，而不是准确的频率。即使如此，这道访题带来的记忆难题也让人气馁，多数受访者的应答可能只是粗略的估计。同样，多数人可能不会仔细考虑下一年的“商业环境”（无论这个概念指什么）。没有理由相信一般受访者会愿意非常用心地回答这道访题，事实上，有充足的证据表明，受访者会有许多办法来简化自己的任务。

这并不是说受访者必须遵循从“理解”开始到“报告”的认知过程。在这些步骤中，某些步骤的反复与其说是例外，不如说是经常发生的情况。另外，虽然大多数人都有问卷调查的经历，也认识到调查的一些特殊约定（如5级量表），但日子还得自己过，日常生活中的问题还需要自己应对。这样，不可避免的是，已形成的习惯与策略会被用来应答访题，并满足调查的要求。

7.2.1 理解

“理解 ”（comprehension）包含了诸如拿到访题、看到提示语、对访题形式赋予意义，以及推断访题含义（如弄清楚访题要了解的信息）等步骤。现在让我们结合NSDUH的某道访题来具体讨论受访者的应答。

> 想一想过去的12个月，从［日期］到今天。我们想知道的是，在过去的12个月，在非医嘱下，有多少天你服用过处方镇静剂，或为了寻求体验或感觉而服用处方镇静剂。［NSDUH］

可以想象的是，受访者的第一个任务是理解这道访题。（虽然访题不是疑问句，却仍传递了对信息的要求，受访者会以对待疑问句的方式来对待这道访题。）正如多数问卷调查的访题一样，这道访题不仅需要某个主题的信息，即处方镇静剂的不当使用；还需要信息的特定形式，即受访者以某种方式使用某种类型镇静剂的天数（1～365）。（在过去一年中，没有不当使用处方镇静剂的受访者不会被问到这道访题，故，答案至少为1。）访题的另一部分（此处没有给出）

进一步明确了应答的形式，向受访者提供了“平均每周的天数”、“平均每月的天数”以及“过去一年的总天数”等多种选择。因此，“理解”一道访题的关键之一（有些人认为应该去掉“之一”）是确定应答模式。问卷调查经常通过提供应答选项或其他应答需要遵循的格式，帮助受访者完成应答。

理解承载了多种活动，诸如分解访题（确定访题的各部分及其相互联系）、对关键词语赋予含义（如访题中的“服用”“处方镇静剂”等词）、推断隐藏于访题的目的，以及确定备选应答之间的界限和潜在的重叠之处等。在完成理解后，受访者可能会超越访题、转向问卷的前一访题或访员或访问环境的提示。尽管NSDUH对访题的设计非常谨慎，尽量避免理解上的含糊，受访者仍然难以理解“处方镇静剂”指称的具体药物，或对“不当使用”的理解会产生分歧。在NCVS中，甚至平常使用的诸如“you”（您，你们——译者注）这类词语也会引起麻烦，“you”到底指你个人？还是指包括你在内的所有家庭成员？（在NCVS中，所有以大写字母表示的单词都是为了提示访员，访题指称的是受访者，与其家庭其他成员无关。）

7.2.2　检索

“检索 ”（retrieval）是指从长期记忆中回忆与访题相关信息的过程。长期记忆指的是储存个人经历的记忆和一般知识的记忆系统，有着巨大的存储能力，可以储存一生的信息。

为了确定NSDUH的题目中哪个可能的应答是正确的，受访者一般都会利用他或她对于该类事件的记忆。对于个人经历记忆的研究并不比

对于人们如何回答调查访题的研究早多少，并且对于个人经历记忆如何组织这一问题，学者也没有一致的意见（参见Barsalou，1988和Conway，1996，二人都曾试图给出相关的研究框架）。然而，对人们如何检索所经历事件的信息这一问题，却取得了某些一致。受访者可能开始时有一些模糊的、包含访题本身的提示（比如，“嗯，在过去一年左右，我曾经在没有处方的情况下数次使用了镇静剂”）。检索线索 （retrieval cue）是对记忆的某个刺激，提供了一些线索来引发对于长期记忆信息的回忆。这个初始的探索也许会唤起长期记忆中某些有用的信息——最好的情况下，会唤起对于访题的精确应答。然而，许多情况下，记忆并不会提供确切的应答，而是以有助于导致某个应答更进一步提示的形式提供相关的信息（比如，“我似乎曾经吃过兰斯（Lance）以前吃过的药，虽然我并不真的需要”）。推断以及基于自身提示（比如，“如果兰斯确有处方的话，大概会包含50片左右的药片”）的记忆将有助于限定应答，并提供更进一步的提示来探索记忆。这个产生提示和检索信息的循环将会继续下去，直到放弃或找到所需要的信息。

一些事物会对检索是否成功产生影响。成功部分依赖于访题涉及事件的性质。某些事件相对于其他事件更难于记忆；当我们试图记忆的事件并不突出，或有许多相同事件，或一开始并未留下什么印象的话，就很难记住。如果你长期以一种自己习惯的方式滥用处方镇静剂的话，就会发现很难记得确切的次数，也很难回忆不当服用镇静剂特定场合的情形。另一方面，如果你仅是一两次不当使用，或就在过去的几天内这样做过，检索就可能产生确切的数字和事件发生的具体环境。

当然，影响某个事件是否容易被记起的另一个重要因素是事件发生的时间。早在一百多年前，心理学家就已经发现事件发生得越早，就越难被记起。来自于NSDUH的那个例子回顾了相对较长的时间段（一年），一些受访者很难记起所有相关的事件，如果有很多相同事件发生，就更难。

影响检索结果的另一个因素是检索线索的数量和种类。NCVS中涉及购物的访题通过向受访者罗列他们可能去过的商店名称（“药店、服装店、杂货店、五金店或便利店”）来帮助他们回忆。这些例子可能会揭示出不同的记忆。检索线索可以提供最多的细节，而当访题中提供的线索与记忆储存的信息不相符时，检索就会失败。举例来说，如果在应答NSDUH访题时受访者并不认为安定是一种“处方镇静剂”，就不会引发相关回忆。

7.2.3 估计和判断

“估计”（estimation）与“判断”（judgment）是整合与补充受访者检索到事物的过程。判断基于检索过程（如，检索的过程是否困难）。另外，判断可以填补回忆的空白，整合检索的结果，或对检索的疏漏作出调整。

虽然NSDUH的访题要求一个确定的数字答案，但提示语却默认受访者回忆的信息可以采用其他形式，比如比率。由于人们通常不会把自己经历某种事件次数的流水账保存下来，故一般不能对NSDUH或NCVS涉及购物的访题检索到一个现成的答案。与之形成对比的是，公司却很有可能对其拥有的雇员保存一份记录，而这正是CES要寻求的关键信

息。CES也证明在某种情况下检索涉及对于实体数据的外部搜索，而不是脑海中对记忆的搜索。通过回忆并计算事件发生次数，受访者可以当场尝试去构建记录。但在事件大量发生的情况下，就很难或不可能全都回忆起来。然而，受访者却较容易根据比率来估计数字。究竟受访者会采取哪种策略——回忆出一份记录，通过记数具体事件来构建记录，还是根据比例来估计数字，抑或去猜——取决于事件发生的次数，调查询问的时间段、某个具体事件的信息是否容易记忆，以及事件发生的频率，所有的这一切都会影响受访者在记忆中储存了何种信息（参见Blair，and Burton，1987；以及Conrad，Brown，and Cashman，1998）。

对态度访题的应答，如消费者调查（SOC）中对商业环境的调查，受访者的反应过程完全不同于基于事实的访题，如NSDUH的处方镇静剂不当使用的例子。也有研究者认为，对态度访题的应答不是预先形成的，而需要受访者将其从记忆中检索出来（参见Wilson，and Hodges，1992）。有多少受访者会留意自己对下一年经济形势的看法，并随着环境的改变和接收到新信息而不断更新自己的看法？以SOC为例，受访者更可能在遇到调查才开始考虑，并与当时想到的和看起来相关的任何事物（如，失业和通货膨胀的趋势，关于世界市场的最新消息，以及最近的股市走势）关联起来形成应答。对态度访题应答时，人们使用的判断策略有可能与应答行为访题时的相同。比方说，受访者在确定下一年经济走势时会试图回忆一些细节，就如人们在应答NSDUH访题时会回忆相关的事件，或根据更一般的信息，类似于NSDUH中的比率以及SOC的长期经济走势。

7.2.4 报告

“报告 ”（reporting）是选择和表达应答的过程，包括使应答与访题的选项相符，以及使当前的应答与之前的应答、自己的感觉或其他标准具有一致性。正如我们已经注意到的那样，NSDUH不仅给出了访题的主题，还给出了应答的格式。可接受的应答可以采用具体天数或比率（每周或每月中的天数）的形式。如果按应答形式划分，则有两种类型的访题。“封闭式”访题向受访者提供了可选择的应答。“开放式”访题则允许受访者用自己的语言应答。当然，在通常情况下，对开放式访题的应答也有一定的约束。大致说来，开放式访题类似于考卷的填空题，封闭式访题则类似于考卷的多项选择题。态度访题总是采用封闭形式，其选项往往会形成一个指标或量表。

受访者选择如何报告他们的应答，将部分取决于他们检索（或估计）的信息与访题要求的一致性。对于要求用数字应答的访题，如NSDUH中的访题，受访者必须让脑海中的判断与给定选项的范围和分布相一致。例如，如果访题的大多数选项表述的是较低的频次，则报告就很有可能向低频次倾斜。或如果没有列出应答选项的话，受访者就得自己确定应答的精确程度并依次给出应答。受访者还可能依据选项出现的顺序（第一个或最后一个选项）和方式（图像或声音）而更关注某些选项。当涉及敏感性访题（如吸毒）时，受访者可能会夸大或隐瞒实际情况，甚至根本不予应答。当采用访员进行调查时，应答压抑感就更容易发生。（参见5.3.5节 ）

7.2.5　反应过程的其他模型

值得指出的是，图7.1描述的调查反应过程模型并不是研究者提出的唯一模型。Cannell，Miller和Oksenberg（1981）提出了一个早期

模型来区分受访者在形成应答时可能采用的两种途径。其中一条途径包含这里讨论的大多数步骤，理解、检索、判断和报告，由此形成准确的、至少是充分的应答。另一条途径则是想走捷径以尽快结束访问或并不想提供准确信息的受访者采取的。这些受访者根据在受访环境中得到的相对表面的提示草率应答，如访员表现或访题暗示的方向。根据这些提示做出应答的受访者有可能受到“默许”（acquiescence，倾向于赞同）或“社会期许”（social desirability，（倾向于通过夸大良好品质和回避不良品质来表现自己的光辉一面）的影响而出现应答误差。

调查反应过程的更近期模型则借用了Cannell模型（Cannell model）中的两条应答途径假设，即由认真的应答者采用的高路径以及由草率的应答者采用的低路径。这就是Krosnick和Alwin（1987）提出的“满足”（Satisficing）模型（参见Krosnick，1991）。按照这个模型，某些受访者是为了“满足”（satisfice）（低路径），而另一些受访者则为了“满意”（Optimize）（高路径）。追求“满足”的受访者并不打算完全搞清楚访题，只要理解到能提供一个差不多应答就可以了；他们也不会去回忆所有相关事件，只要回忆出的事件足够产生应答就行了；诸如此类。“满足”类似于Cannell两条路径模型中较为草率的分支。类似的，追求“满意”的受访者遵循的则是较为认真的分支。在其后的著作中，Krosnick还区分了追求“满足”的受访者为了尽快应答所采用的特定策略。例如，他们可能会在回答“同意—不同意”态度访题中都应答“同意”，即“默许”应答策略。

对于应答策略的评论

受访者可以采用多种策略应答。其中的一些策略如选择“不知道”或“没有想法”或每题选择相同的选项，可以极大地减少应答所花费的精力。这就是调查中常见的“满足”策略，受访者尽可能不费力地完成应答要求。

和Cannell模型一样，Krosnick的满足理论对于持续变化的各过程之间作了明显区分。受访者可能会以不同的认真程度来处理不同的访题，同样，也可能对反应过程的各部分予以不同的关注。受访者不用心听取访题并不意味着他或她不会努力检索。出于种种原因，受访者会认真或草率地处理各个认知步骤。我们不妨认为Cannell和Krosnick区分的两条途径是受访者以不同深度考虑访题并形成应答这一连续过程的两个极端。

对于调查访问，还可以通过进一步的研究来寻找更适切的模型。其中一项有重大意义的课题就是探索在调查互动中作为中介的计算机的角色。如CAPI或ACASI中的计算机究竟是调查互动过程中的另一个要素呢，还是和静态纸质问卷的作用一样？计算机辅助在面访中对受访者和访员的作用是什么？计算机程序设计能怎样改变受访者的行为以得到更高质量的数据？

7.3　应答调查访题中的问题

有一个应答过程模型（即便是像图7.1中那样相对简单的模型）的好处之一是让访题设计者能系统地考虑可能出现的问题以及影响应答不准确的因素。正如我们知道的那样，调查的目的是降低误差，而误

差的重要来源之一就是测量误差，即问题的真正答案和最终出现在数据库中的答案不同。（虽然这里对测量误差的定义不大适合于态度访题，我们仍然希望对态度访题的应答与想要查明的真正态度是相关的。因此，我们选择了与受访者态度非常相关的态度测量。）

对应答认知分析的重要假设是，在产生答案认知中的错误导致了应答误差。这里，我们列出了应答过程中引起调查误差的七种错误：

Fowler论访题中的含糊措辞

1992年，Fowler通过研究指出，在前测中去掉含糊措辞会影响调查应答。

研究设计：对有60道访题的问卷进行了100人的前测访问，并对访问进行了录音。通过行为编码记录了每次访问中访员和受访者的问答顺序。获得了7个有待澄清的含糊措辞。这7个含糊措辞导致了至少15%的受访者应答不当。访题的修正版本试图去掉含糊措辞。第二轮前测访问了150人。然后，对两次前测的应答分布和行为编码数据进行了比较。例如，第一轮前测包含了以下问题：“你每周平均有多少天会食用黄油？”第二轮则处理了有二义性的措辞“黄油”，将访题转变为：“下一个问题仅涉及黄油，而不涉及人造黄油。你每周平均有多少天会食用黄油？”

研究发现：相对第一轮前测，受访者要求对访题进行解释的情形以及应答不当均有下降。应答的分布改变了；比如，此题中：

从不吃黄油

第一轮前测	33%
第二轮前测	55%

研究者发现，排除掉人造黄油使得报告“从不吃黄油”的人增加了。

研究局限：对应答分布的真实值缺乏外部标准作为参考。其结果无法判明行为编码产生的错误需要对访题措辞做多大程度的改进。研究认为，访题措辞应保持一致。

研究意义：研究证明有行为编码的前测可以辨别出有错误的访题。通过行为编码显示了访题措辞的变化可以改进访问中的交流互动，进而影响调查结果。

1）不能编码搜寻到的信息。

2）对于访题的误解。

3）遗忘或其他记忆上的问题。

4）错误的判断或估计方法。

5）形成答案时的问题。

6）或多或少的有意误报。

7）不遵循提示语。

有人曾列出了更多、更详尽的应答错误（如Sudman，Bradburn，and Schwarz，1996；Tourangeau，Rips，and Rasinski，2000）。这些方法都有一个共同的假设，即测量误差一般可以追溯到应答过程的问题（如，受访者根本没有所需的信息，或忘记了，或曲解了题意，或作出了错误的判断，等等）。

7.3.1　编码问题

一个人经历某件事，并不意味着他一定记住许多信息。对目击者证言的调查发现，目击者常常漏掉事件中的重要细节（参见Wells，1993）。问卷调查询问的是更普通的经历，受访者可能记住的信息就更少了。因此，受访者的事后估计，大体基于通常发生的情况。由A. F. Smith的研究证明了这一点。他把受访者对一日三餐的事后报告与其记录的食物日记进行了比较，报告与日记之间的差距如此巨大以至于Smith（1991，第11页）断定，“受访者对饮食的报告……大部分是对吃过什么的猜测。”大多数人都不曾关注自己吃过什么，所以，询问受访者吃过什么，得到的报告就不会准确。

从这个例子中我们可以学到一些东西。人们不可能提供他们没有的信息；如果人们不从一开始就对信息进行编码，那么，无论设计多么精妙的访题都不可能引导出准确的应答。前测的重要目的之一就是确定受访者拥有调查需要的信息。

7.3.2　误解访题

即便受访者知道访题的答案，如果他们误解了题意，也不可能提供正确的应答。虽然很难描述误解题意的发生状况，但有迹象表明，这种情况经常发生。

证据之一是Belson（1981；同时参见Belson，1986）被广泛引用的研究。在研究中，Belson要求受访者自己描述访题中关键词的词义。他发现甚至像“you”或“周末”这样直白的词语，受访者都有不同的理解。（类似于“you”指的仅仅是你自己？还是包括你的配偶或家庭？星期五算不算“周末”的一部分？）Belson还研究了一个在今天看来非常明显的例子：

> 你认为孩子观看除西部片之外的暴力节目会对他们有不良影响么？

Belson的受访者对“孩子”给出了数种解释。“孩子”有两种基本的词义：年幼的人，不管他和你的关系如何；或你的后代，不管他年龄有多大。在第一种定义中，究竟多大年龄以下的人可被称作“孩子”会随着具体环境的变化而变化（如发型、看限制级电影的年龄以及购买酒精饮料的年龄等）。Belson发现，在受访者中确实出现了类似的对“孩子”定义上的变化。他也发现一些特殊定义（如，紧张不安的孩子，某个人的孙子等）。如果“孩子”被给予了多种解释，那么像“不良影响”之类事先故意弄含糊的措辞就更有可能有多种解读了（同时也并不清楚为什么这道访题不包括西部片的暴力，或受访者将会如何理解）。

如果受访者不愿询问某些措辞的含义或不愿承认自己看不懂访题，那么，理解访题的错误可能会导致错误应答。一些研究曾向受访

者询问虚构的问题（如“公共事务法案”）时发现，仍有40%的受访者愿意对这类问题置评（Bishop，Oldendick，and Tuchfarber，1986）。在日常生活中，一个人向另一个人提问的前提是，提问者认为被问者很可能知道问题的答案或至少值得向被问者提问。同样，受访者会认为自己应该知道诸如“公共事务法案”之类的问题或问题中使用的词语。当他们遇到理解困难时，就会觉得向别人求教很丢人，进而胡乱应付。另外，访员可能受到不要向受访者解释访题的训练或仅仅会提供无足轻重的回应（如再念一次原来的访题）。不幸的是，正如Belson指出的那样，即便对日常用语，人们也常常有不同的理解；当访题中包括生僻词或技术术语的话，人们的不同理解就更多了。

Tourangeau，Rips和Rasinski（2000）区分了调查中可能出现的7种理解困难：

1）语法不明确。

2）过分复杂。

3）错误预设。

4）模糊概念。

5）模糊数量词。

6）生僻词。

7）错误推断。

前三种与访题的语法形式有关。“语法不明确 ”（grammatical ambiguity）指的是访题有两种或两种以上的解释。比如，即便是简单如“你是来访消防员么？”这样的句子也会有两种不同的解释：你是来访的消防员，还是你打算去拜访消防员？在现实生活中，具体的语境可以帮助确定访题的含义。但在调查中，语法不确定会在不同受访者中导致不同的理解。在调查中更常见的是语法过分复杂。下面是由Fowler（1992）讨论过的一个例子：

> 在从1987年1月开始的过去12个月中，您曾经有多少次因健康问题去看医生或其助理或与之交谈？不计算您住院时发生的此类情况，在其他任何情况下去看任何科的医生都应该被包括在内。

“过分复杂 ”（excessive complexity）指访题的语法结构妨碍了受访者推断其实际含义。上例中的不当之处就是列出了多种可能（看医生、与助理交谈等等），访题最后的提示语进一步增加了复杂性。类似于上例的问题在于，受访者不可能将访题中的所有可能与要求记住，从而导致忽略其中的部分内容。

“错误预设 ”（faulty presupposition）指访题假设了并不真实的情况，访题没有意义或不能应用于现实。比如，假设询问受访者是否赞同以下说法，“家庭生活总是不如意的，因为男性过于关注自己的工作。”这道访题预设了男性过于关注自己的工作；不赞成预设的受访者（如，认为大多数男性都很懒惰的人）并不能对访题提供有意义的应答。其实，所有访题像一幅预设了部分事物的图画，而要求受访者填上这幅图画的空白部分。

随后的三种错误与访题措辞的意义相关。正如Belson所指出的那样，很多日常用语都是很模糊的，不同的受访者可能有不同的理解。因此，访题的表述，越具体越好。比如，针对儿童的访题应说明具体年龄范围。问题是，要把一个模糊概念 （vague concepts）解释清楚，就会增加访题的复杂性。上例访题就是此种情况。这道访题试图定义清楚“医生门诊”概念。一些访题还在应答选项中使用了相对模糊的语言（“有点不同意”、“经常”）。不幸的是，受访者之间可能对于怎样的频率才应算“经常”并没有一致的意见，如此，不同的受访者会以不同的方式处理选项。另一种理解上的困难来源于受访者不知道陌生措辞 （unfamiliar term）的含义。编问卷的人常常是问卷主题的专家，很有可能高估受访者对某些措辞的熟悉程度。想要知道人们养老金计划的经济学家可能会不自觉地使用诸如“401（k）”或“SRA”这样的术语而不作定义。不幸的是，这些术语很可能会使一些受访者感到迷惑。

也有研究发现受访者可能会过度理解访题，从而对其目的作出错误推断 （false inference）。考虑一下来自综合社会调查（general social survey）的这道访题：

您是否想到某种情境允许警察殴打成年男性公民？

人们很容易想到一些回答“是”的情境（看看警匪片就知道了），但仍有受访者（大概30%）回答了“否”。显然，一些受访者没有按照字面去理解这道访题。取而代之的是，他们按照自己感知的调查目的（评价警察的暴力行为）作出了应答。对调查目的的推断，是理解过程的自然组成部分，却会使受访者误入歧途。例如，一些研究

表明，如果访题问及“总的来说，您觉得最近过得如何？您是觉得非常快乐，有些快乐，还是并不快乐”，且位于一道婚姻幸福之类的访题之后，那么，在这道访题的应答中，受访者就不会考虑自己的婚姻是否幸福（Schwarz，Strack，and Mai，1991）。在日常谈话中，每次我们开口，都会被预期要讲新的东西；这种预期导致受访者认为这道一般性的访题是没有被询问过的，即除了婚姻生活外的其他生活的。不幸的是，如此推断可能与调查设计者的愿望并不相符。

这一部分强调了语言在理解阶段的重要性。在许多情况下，清晰地定义访题的术语（为了消除含糊性）和增加受访者理解题意的负担之间存在着矛盾。我们需要更多的研究来确定访题的表述应清晰到何种程度，何时向受访者提供所需的解释，以及访员和受访者间的互动交流如何影响后者对于访题语言的理解。

7.3.3　遗忘和其他记忆问题

调查误差的另一个潜在来源是遗忘相关的信息。受访者有时会完全不能记起有关事件，有时会大概或不确切地记起某事。回忆失败的种种情形对应答有着不同的影响，对其进行区分是有益的。

1）访题的措辞与初始记忆编码不同。

2）随时间流逝对事件印象扭曲。

3）检索失败。

4）重构误差。

第一种回忆失败发生在受访者用来记忆的编码措辞与访题措辞显著不同之时，由此导致访题不能唤醒所需的记忆。例如，受访者也许认为餐饮葡萄酒并不属于“酒精饮料”。在这种情况下，询问每周酒精饮料消费的访题就不能唤起相关的记忆。同样，大多数人都不会把去五金商店买东西算作“购物”；因此，在NCVS访题中就明确地列出了“五金商店”。（“在过去的6个月中，即从________开始，平均而言，您购物的频率？譬如去药店、服装店、杂货店、五金店或便利店。”）在研究者设计问卷时，常常会召开焦点小组会来了解潜在的受访者如何思考和谈论调查主题。当问卷措辞与受访者的记忆编码措辞匹配时，理解和检索过程都会得到改进。

记忆不准确的第二个来源是随时间的流逝对事件细节的模糊。当向不在场的其他人叙述、与有相同经历的人一起回忆，或事后思考时，大多数自我经历记忆都可能混杂着初始信息即事件发生时或紧接事件发生后的信息，以及事件发生后添加的信息。例如，当回想自己高中毕业典礼时，记忆中会包含当时的信息以及翻看毕业纪念册的照片或录像时得到的信息。这就是记忆研究学者所称“复述”（rehearsal）在保持记忆的生动形象方面扮演的重要角色（参见Pillemer，1984；Rubin and Kozin，1984）。不幸的是，我们很难确定记忆信息的来源，也并不总能对直接经历的事和听说的、事后推断的事进行区分。因此，我们回想个人或集体经历时带来的种种扭曲和修饰不可能与初始记忆信息相区分。这种“事后信息”不一定不准确，却有不准确的可能，且一旦掺杂对某事的记忆，就很难把它去除。

回忆困难的另一个来源是检索失败，即不能把长期记忆的信息回忆起来。我们已经注意到检索失败（retrieval failure）的一种原

因，即访题不能唤起对某事的回忆，因为访题措辞与记忆初始编码措辞不同。检索失败的另一个原因是遗忘相似经历中的某个经历。随着时间的流逝，人们会越来越难记住事件的细节，比如说把某次医生的出诊从类似事件中区分开来；同时，相类似事件如医生出诊、购物、出差等会被模糊成“一般记忆 ”（generic memory）（参见Barsalou，1988；Linton，1982）。随时间推移，相同事物的累积意味着我们越来越难以记起某些特定事件。时间流逝的影响很可能是一百多年来遗忘研究中最有力和最可靠的发现（Rubin & Wetzel，1996）。虽然研究者们仍不清楚遗忘与时间流逝的具体函数关系，但已经清楚的一点是开始时遗忘的速度很快，之后就慢了下来。在给定时期内，遗忘量还取决于讨论事件的性质。一项研究发现，人们在50年后仍能记住近一半同学的名字（Bahrick，Bahrick，and Wittlinger，1975）。图7.2显示了准确记忆某些事物的百分比。虽然同学很容易被记起，但随时间推移，分数的被遗忘率上升很快。在调查中，对检索失败进行修正的最佳方法是提供更多的检索提示和允许花更多时间来回忆。表7.1（修改自Tourangeau，Rips，and Rasinski，2000）提供了影响回忆更综合的因素及其对调查设计的意义。简而言之，容易被记起的通常是最近发生的、独特的、发生于其他容易被记起事件附近的事物，以及在受访者一生中重要的事件。同时，有充足的相关提示、给予受访者充分时间并鼓励其认真思考的访题会有较好的效果。

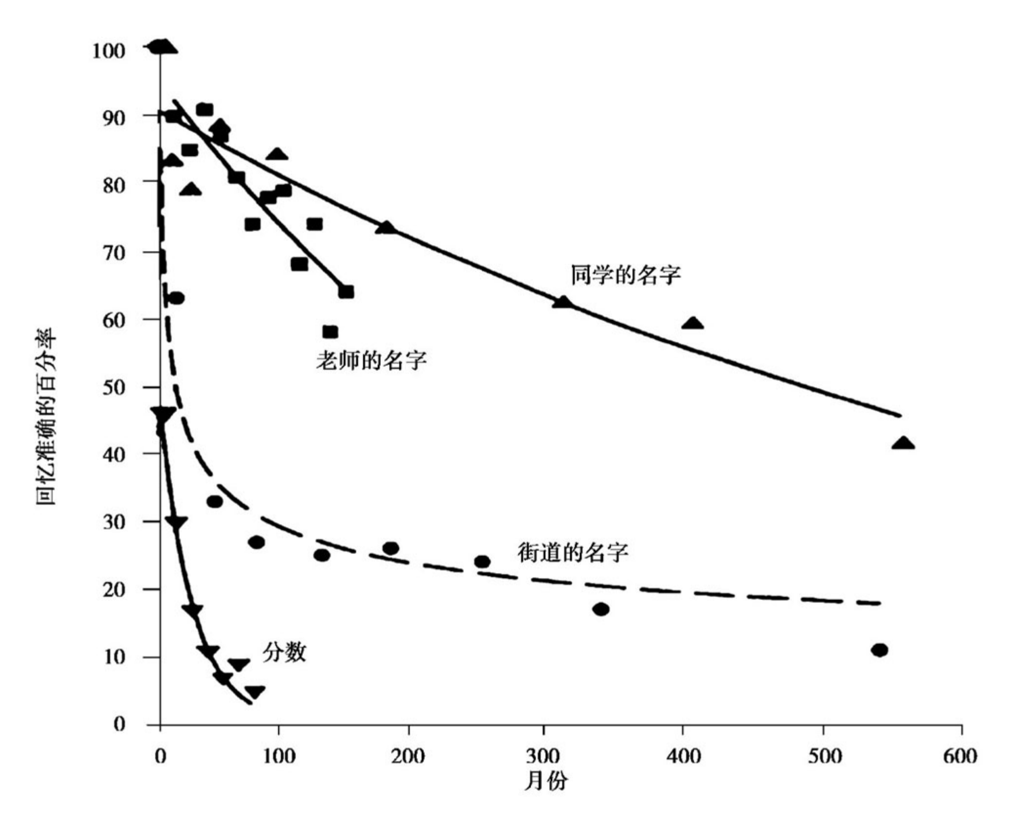


图7.2　不同个人信息类型回忆的准确性（数据来源：Tourangeau，Rips，and Rasinski，2000）

表7.1　影响回忆因素的总结

变　量	发　现	对于调查设计的意义
有关事件的属性		
发生的时间	较长时间以前发生的事件比较难回忆起来	缩短参照期
接近时期界限	在重要时期界限附近发生的事件比较容易记起	使用事件史研究等方法以改善回忆
独特性	独特的事件容易被记起	按照目标事件的性质来设计参照期；使用多种提示来启发对个别事件的回忆
重要性，情感影响	重要、涉及情感的事件容易被记起	按照目标事件的性质来设计参照期；
访题特征		
回忆顺序	追溯性回忆效果较好	不清楚是否在调查中追溯性回忆也较好
提示的数量和类型	多个提示强于单个提示；事件类型（“是什么”）的提示优于参与者或地点（“是谁”或“在哪里”）提示，优于时间（“什么时候”）提示	提供多个提示；使用分解法
完成应答的时间	时间充足则回忆效果较好	使用较长的访题引导语；减慢访问的节奏

回忆及报告的准确性随时间推移而下降这一趋势产生了一个重要的测量误差模型。在模型中，μ_i 代表的是第i 个受访者经历的访题询问事件的数量。即，如果没有回忆困难的话，第i 个受访者对访题的应答是μ_i 。而在实际调查中，第i 个受访者报告不是μ_i ，而是y_i ：

$$y_i = \mu_i \left(ae^{-bi} + \varepsilon_i \right)$$

式中，a 代表的是在考虑敏感性或社会期许影响后事件被报告的比例；b 代表的是作为时间函数的下降率；而ε_i 则是模型中第i 个受访者产生的误差；e即欧拉常数，自然对数的底。模型刻画的是事件被准确报告的比率呈指数下降（在访问前不久发生的事件下降速度很快，在较早前发生的事件则下降速度减缓）。有关文献显示了对独特的、发生于其他被容易记起事件附近的以及对受访者一生重要的事件

而言，*a* 接近于1.0，*b* 而接近于0.0。对容易被遗忘的、不引人注意的事件而言，*a* 也可能接近于1.0，*b* 值却很大。从图7.2可以看出，指数下降模型相对于其他模型来说，更拟合经验数据。

记忆失败的最后一个来源是重构 （reconstruction）或填充不完整记忆的努力。重构通常基于经常发生或当下发生的情况。例如，Smith对饮食的研究发现，受访者用来填补饮食记忆空白的猜测依据的是常吃的食物。Bem和McConnell（1974）揭示了一种不同的策略。在他们的研究中，受访者用他们当下对某事的观点去推测自己以前的观点。这种“回顾性”误差重复出现的数次（参见T. W. Smith，1984）。当下的状态也会影响对以前其他事物的记忆，如痛苦、酗酒史或吸毒史、去年的收入等（更多的例子请参见Pearson，Ross，and Dawes，1992）。人们常常通过审视当下并向前推测来重构过去，且暗自假定访题涉及的性质和行为是一成不变的。另一方面，当我们记起确实发生过某种变化时，又会夸大这种变化。

许多调查项目要求受访者报告在某一确定时间段，即参照期（reference period）发生的事件。前三个调查范例都指定了参照期，即从受访者接受调查的时刻回溯到过去的某个时点（如过去整6个月的那一天）。这些访题都假定受访者可以相当准确地回忆事件发生的时间。不幸的是，日期可能是事件中最难被人准确记住的属性（参见Wagenaar，1986）。虽然人们能准确记住某些事件（如生日，婚礼或诸如此类）的日期，事实上，人们却一般不会注意或记忆日期。由于确定事件发生的日期有困难，受访者可能错误地报告参照期之前发生的事件，从而犯下远溯 （telescoping）的错误。远溯指把很久之前发生的事件当作最近发生的事件。实际上，最近的研究表明，向前

远溯也很常见。随着时间流逝，在回忆事件发生日期时，人们会在前后两个方向上犯较大的错误（Rubin and Baddeley，1989）。

尽管如此，从整体上来说，远溯还是会导致多报。例如，在一项经典研究中（参见文本框 ），受访者对住宅维修的应答有40%的远溯（Neter and Waksberg，1964）。有时，人们会运用“约束”手段来减少历时调查的远溯误差。在受约束的访谈中，访员会和受访者一同回忆受访者在前一次调查中的应答。NCVS就运用了这一步骤以排除前一次调查已经报告过的事件。在七轮NCVS调查中，第一次调查询问了在过去6个月的刑事犯罪受害事件，不过此次调查的数据并不会用作估计的参数。第二次调查则使用了第一次调查获得的数据来“约束”受访者。访员通过查询首次访谈的报告来确认第二次调查报告的事件与第一次调查报告的事件是否有重合。以后几轮的调查，也采用相同的步骤，即用前一次调查的数据作为“约束”。这种方法极大地降低了由远溯导致的受访者重复报告的可能性。

Neter和Waksberg论报告误差

1964年，Neter和Waksberg发表了对不同设计报告误差的研究结果。

研究设计：两种调查设计特征具有系统差异性，包括访谈是否受到约束（受到约束即向受访者提醒前次调查的结果）以及需要受访者回忆时期的长度（如1个月，3个月或6个月）。研究的情景是，向家庭成员询问住宅维修的次数及每次的花费。

研究发现：相对于约束访谈，无约束访谈报告了相当高的花费金额（高出55%）。大规模维修的增长较高。研究者认为，受访者将参照期以前发生的事件纳入进来了，对较少发生事件的远溯较严重。与回顾过去1个月的情况比较，则回顾过去6个月报告的次数要少，即在较长的参照期中不成比例地漏掉了较小的维修工作。6个月参照期报告的每月小型维修工作要比1个月的少32%。研究者认为，这是报告失败与追溯的混合效应。

研究局限：此类工作和花费并没有其他独立的数据。因此，研究者假设了受约束访问提供了最佳报告，并以此获得结论。但一些受访者可能被询问了多次，且每次的报告数都不相同。

研究影响：这一研究有效地提醒调查设计者通过关注参照期的长度来改善调查质量，并鼓励运用受约束访谈，如全国刑事犯罪受害者调查。

7.3.4　行为访题的估计过程

根据回忆，受访者也许会被迫对行为发生的频率进行估计或对态度进行判断。考虑一下本章的两个例子：

> 如果把国家商业环境当作一个整体，您认为在接下来的12个月，将会遇到好的经济时期？不好的经济时期？还是其他的什么？［SOC］

在过去的12个月，即从________［日期］到今天。在非医嘱条件下或为了寻求体验或感觉，有多少天您服用过处方镇静剂？［NSDUH］

有些受访者可能对SOC询问的经济形势预先作过判断，但大多数受访者则会临时拼凑应答。同样，仅有少数受访者会保存其不当使用处方镇静剂的记录；多数人最多只能对整体状况作出估计。（请注意，在NSDUH中，允许受访者选择不同的应答模式是有帮助的，譬如平均每周几天，每月几天，或总共多少天。）在态度和行为访题中，对受访者现场作答的要求可能会导致应答中的误差。

让我们先来看类似NSDUH中的涉及行为发生频率的访题。除了回忆事件发生的准确时间外，受访者还可以采取以下三种估计策略应答：

1）回忆具体事件，然后加总；在考虑遗忘的情况下，应答者会把答案调整得稍大一些（“回忆并计算”，recall-and-count）。

2）回忆事件发生的大致频率，然后依据参照期的长度进行估计（“基于频率的估计”，rate-based estimation）。

3）从大致的印象推断出一个数字（“基于印象的估计”，impression-based estimation）。

例如，一位受访者也许回忆有三次不当使用处方镇静剂，并报告“3”作为应答。另一位受访者可能记起在过去一年大概每月有一次不当使用处方镇静剂，并且报告“12”作为应答。第三位受访者的印象

中“有几次”不当使用处方镇静剂，并且报告“5”作为应答。显然，这些应答采用的不同策略会产生不同的误差。

“回忆并计算”策略易于因遗忘而遗漏，也会因远溯而误报。这两种误差来源之间的权衡决定了受访者可能有规律地过高或过低报告发生的事件。一般来说，要报告的事件越多，基于这一策略应答的准确性就越低；当事件较多时，受访者会较难全部记起或在头脑中加总。因此，当事件发生次数较多或参照期较长时，受访者会转而使用其他策略（Blair and Burton，1987；Burton and Blair，1991）。

一些研究探讨了受访者在遇到如NSDUH的行为频率访题时的应答策略，发现当需要回忆的事件数量增长时，受访者运用“回忆并计算”策略的受访者人数迅速下降。同时，在有7个左右或7个以上的事件需要报告的情况下，受访者会转而使用“基于频率的估计”策略。研究文献显示“基于频率的估计”会导致对行为发生频率的过高估计。显然，当频率存在波动或有例外时，人们会过高地估计行为发生的频率。

对行为发生频率访题而言，最容易出现误差的策略还是“基于印象的估计”。当采用封闭式访题时，访题列出的选项会影响基于印象的估计。当应答选项集中于较低频率时，应答会倾向于较低的频率；当应答选项集中于较高频率时，应答也会倾向于较高的频率。此处文本框 的内容显示了一项研究的结果，这项研究调查了人们平均每天有多长时间在看电视。依据受访者遇到的不同选项，分别有16.2%和37.5%的受访者说自己每天看电视的时间多于2.5小时。当采用开放式访题时，基于印象的估计又可能得到野值。

过低或过高的报告

当受访者报告了实际上并没有发生的事件或报告了比实际情况更多的事件，这就是“过高的报告”。某些类型的事件通常会被过高地报告。例如，在任何一次选举中，有比实际投票人数更多的人会报告他们投了票。与之相反的一种错误叫作“过低的报告”，即报告了比实际情况更少的事件。

7.3.5　态度访题的判断过程

态度访题的应答涉及与行为访题不同的认知过程，不过，仍然有学者（如，Tourangeau，Rips，and Rasinski，2000）认为两种应答之间的共同点要大于不同点。对频率进行推测基于四类信息：确切的记录，印象，一般信息，以及特定记忆；在态度访题应答中，也有相应的类似信息。例如，有些受访者（如关注股市变化的经济学家）可能对经济形势有清晰的观点，可以应答SOC访题。（“如果把国家商业环境当作一个整体，您认为在接下来的12个月，将会遇到好的经济时期？不好的经济时期？还是其他的什么？”）另一些受访者可能仅有模糊印象。（“呃，我有天读了华尔街日报的一些东西，看起来好像未来经济走势并不好”）。正如人们对做过的事只有模糊印象一样，当要求对某事或某物评价时，我们一样，也只有模糊印象。或者，由于缺乏已有的判断（即便是一个大致的判断），我们要么从上到下即依靠更一般的价值观或倾向性来构造判断；要么从下往上即通过访题

的某个特定观点来构造判断。后面的两个策略类似于回答行为访题时的一般信息或回忆并计算策略。

Schwarz等人论量表效应

1985年，Schwarz，Hippler，Deutsch，和Strack发表了多项研究的结果，以讨论对量表应答效应的测量。

研究设计：把两个不同的随机实验嵌入到更大的调查中，把一种形式的设计用于一半样本，另一种形式的设计用于另一半样本。其中，一项实验的样本量是132名成人；另一项实验的样本是79名办公室职员。针对一组询问看电视时间，有6个选项，中间选项为1～1.5小时和1.5～2小时；另一组的中间选项为3～3.5和3.5～4小时。

较低选项		较高选项	
应答	%	应答	%
<0.5 小时	7.4	<2.5 小时	62.5
0.5~1 小时	17.7	2.5~3 小时	23.4
1~1.5 小时	26.5	3~3.5 小时	7.8
1.5~2 小时	14.7	3.5~4 小时	4.7
2~2.5 小时	17.7	4~4.5 小时	1.6
>2.5 小时	16.2	>4.5 小时	0.0
合计	100.0		100.0

研究发现：与拿到“较高选项”组比较，拿到“较低选项”的受访者倾向于报告较少的看电视时间。例如，在较低选项组中有16.2%的人报告每天看电视的时间多于2.5小时，而在较高选项组中，多于2.5小时的有37.5%。

研究局限：其中一项研究使用了限额样本，另一项调查的是办公室白领，限制了研究结果的推广。

研究影响：证明了应答量表可以影响对于行为的报告。研究者现在尽量采用接近于人群均值的中间项。

当没有可运用的事先判断时，受访者对态度访题的应答会极大地受到访题措辞以及调查环境的影响。看看下面两题。这两题都是在20世纪50年代早期用来测量对朝鲜战争支持度的。

> 您认为美国在进行朝鲜战争时犯了错误吗？还是没犯错误？［Gallup］
>
> 您认为美国派遣美国士兵参与朝鲜战争是正确的还是错误的？［NORC］

与Gallup的访题比较，NORC的访题表现了更多对朝鲜战争的支持。在其后的一系列的实验中，Schuman and Presser（1981）发现加上短语“为防止共产主义的占领”后，公众对于美国军事干预的支持上升了15个百分点。一些研究也发现了类似的文字措辞的影响：人们对用于制止犯罪、救济穷人、戒毒的经费支持要多于对法律执行、社

会福利、戒毒后身体康复经费的支持，以及诸如此类的情况（Rasinsiki，1989；T. W. Smith，1987）。访题措辞可以帮助（或影响）需要从一般价值观来形成自己对某个特定事件观点的受访者。NORC的措辞显然提醒了某些受访者较大的问题：共产主义的扩散，从而帮助受访者形成了对美国在朝鲜所扮演角色的判断。

调查环境也会对受访者评价某个问题产生影响。大多数人对于态度的判断都来自于比较。当我们评价某个政治人物时，几乎不可避免地涉及比较，与其竞争对手的比较，与出色政治人物的比较，与典型政客定义的比较。比较的标准会直接影响到对政治人物的判断。如果用里根政府作为比较标准，则某个民主党人可能认为克林顿在任美国总统期间的表现很出色；如果用罗斯福政府作为比较标准，则克林顿的表现可能就不那么出色了。

7.3.6　形成应答

一旦受访者形成估计或初始判断，接下来的问题是，如何将判断转化为可接受的形式。调查访题有多种形式，我们集中讨论如下最常见的三种：

1）要求数字答案的开放式访题。

2）顺序选项的封闭式访题。

3）分类选项的封闭式访题。

下文以BRFSS的访题为例来讨论。

1）现在，考虑一下包括疾病和受伤在内的您的身体健康状况，在过去 30天中，您有多少天身体不佳？

2）您认为大体上您的健康状况是：

1. 极好

2. 非常好

3. 好

4. 一般

5. 不好

3）您是：

1. 已结婚

2. 已离婚

3. 配偶去世

4. 分居

5. 从未结婚

6. 未婚同居

对第2、3道访题，访员要“读出”访题的选项（不读出选项附带的数字）。在BRFSS中，几乎所有访题都是这三种形式中的一种，或“是、否”类型。“是、否”类型接近于封闭分类选项访题。事实上，不单是BRFSS的访题采用这些形式，大多数其他调查也这样。

这三种形式的访题对受访者分别提出了不同的挑战。在数值型开放式访题中，受访者在将模糊的潜在判断（“我这个月过得很不好”）转化为确切的数值（“我病了3天”）上存在很大的困难。在调查中，开放式访题的流行是有原因的。从理论上来说，开放式访题相对于封闭式访题会产生更准确的应答。即便是应答选项的分类很细，当应答被分类时，不可避免地会损失信息。此外，应答选项还有上限或下限。例如，BRFSS中有一道开放式访题问及受访者在过去的12个月中有多少性伴侣。如果封闭式访题提供极限选项（如，“10个或更多”），就会使人群中在性方面较为活跃的那部分人产生不准确的信息。因此，相对于开放式访题，封闭式访题由于分组和取值的限制而失去了部分信息。

然而，在现实中，受访者似乎认为开放式访题不是真在寻求准确的数值。例如，Tourangeau和他的同事发现，在性行为调查中选择“10个或更多”的人，大多总选择5的倍数作为应答（Tourangeau，Rasinski，Jobe，Smith，and Pratt，W.，1997）。受访者还会以取整的方式报告其他的值，如在照料有残疾的亲戚时感受到的压力强度（Schaeffer and Bradburn，1989）或多久前完成了上一次调查（Huttenlocher，Hedges，and Bradburn，1990）。当题目问及百分率时，受访者提供的应答倾向于聚集在0，50，100附近。导致对应答取整的有多种因素（如搞不清实际数量或对应报告的数值感到尴尬），其中最关键的因素还是许多受访者在把估计和判断转化为数值

上有困难。受访者会从限定的值域中选择一个值来简化任务，之后把选择的值域作为取整值报告。

第二题，运用量表选项也有特定的困难。对某些类型评价（如自我评定），受访者会不好意思报告负面选项，从而产生了“正面偏差”（positive bias）。对其他类型的评价，受访者会倾向于避免极端选项。特别是当选项前有数标时，也会影响应答。Schwarz及其同事（Schwarz，Knäuper，Hippler，Noelle-Neumann，and Clark，1991）要求受访者评价自己一生取得的成功。其中，一组受访者使用从-5到5的量表；另一组受访者则是从0到10的量表。在两种情况下，量表正负选项有相同的文字内容。对两种数标来说，评价都倾向于落在正面选项上（整体的正面偏差），但使用数标为-5到5的评价，聚集于正面选项的现象更为明显。按照Schwarz和其同事的观点，负数与从零开始的数，表达的是不同意思。从零开始的数表达的是在一生中缺乏成功，负数代表的却是失败。

评价中至少还有两种因素会对应答有影响，语标和选项的数量（例如5级、9级）。Krosnick and Berent（1993）进行了一系列研究来比较两种类型的应答量表，其中一种仅用语言标注两端的选项，另一种则对所有选项都用语言进行标注。另外，他们还比较了典型的量表题（一次性显示所有选项）和嵌套访题（开始提供某些选择，如“您是共和党人，民主党人还是无党派人士？”，之后，则是更为细致的选项，如“您是非常坚定的，还是较一般的民主党人？”）语标和两步走嵌套题都提高了应答的可信性。Krosnick and Berent认为，语标有助于说明选项的含义，而嵌套结构则将受访者的任务分解为更简单的判断，从而降低了应答的难度。除了语标和嵌套结构以外，选项的数量也会影响访题的难度。当选项太少时，评价尺度难以用来区

分受访者的不同判断；当选项太多时，受访者又不能有效地区分相邻的选项。Krosnick and Fabrigar（1997）认为，七项选项是最佳的选择。

最后一种常见访题形式是向受访者提供无顺序的选项（正如第3题给出的婚姻状态）。这种形式的困难之一是，受访者不会等到听完所有选项就会选择听到的第一个合理应答。例如，对BRFSS中的婚姻状况访题而言，受访者会选择“从未结婚”，而没有认识到最后一个选项“未婚同居”或许更准确地描述了自己的状态。若干项研究比较颠倒了选项顺序的影响。这些研究发现了两种类型的影响：首呈效应和近呈效应。“*首呈效应* ”（primacy effect）指被首先列出的选项（或接近于首项的选项）有更大的被选择机会。“*近呈效应* ”（recency effect）则恰恰相反，尾项或接近尾项的选项更容易被选择。多数研究者认为，受访者分别考虑各个选项并且会选择首先碰到的、提供合理应答的选项。用Krosnick（1999）的话来说，受访者是为了满足，而不是为了满意；选择的是足够好的答案，而不是最好的答案。人们走捷径的趋势解释了为什么首呈效应很常见。之所以发生近呈效应，则是由于受访者不一定按照问卷给定的顺序阅读选项。当访员向受访者读出访题时，访员读出的最后一个选项也许是受访者开始思考的第一个选项。（与之形成对比的是，当受访者自己阅读访题时，他们更有可能按问卷给出的顺序阅读和考虑选项。）由于这些差异，电话调查中的受访者容易受到近位效应的影响，邮件调查中的受访者则容易受到首位效应的影响。选项顺序的作用并不只是调查中才有的现象。Krosnick指出，在政府官员的选举中，候选人的出场顺序也会影响到每个人得到的选票。

7.3.7 有意误报

到目前为止，我们一直讨论的是受访者处理由访题带来认知困难的努力，可浏览一下汗牛充栋的调查问卷就会发现另一种难题。例如从NSDUH中摘出访题：

（a）在过去的30天中，您有多少天吸食可卡因？

（b）在过去的12个月中，您有多少次酗酒经历？

（c）您最近一次抽烟是什么时候？

人们很容易想象，在理解第一题或检索并形成应答时不会有什么困难，可受访者依然会给出错误应答。调查研究者将此类访题称为“敏感性”或“威胁性”访题。在试图监控毒品非法使用以及因吸毒引起的艾滋病传播时，国家调查中的这类访题很常见。“敏感性访题”（sensitive question）是让受访者觉得被冒犯或尴尬的访题。例如，个人收入或性行为的访题就属此类。对此类访题，受访者极有可能拒答或故意错误应答。敏感性访题让受访者进退两难，一方面受访者答应了要向研究者提供信息；另一方面又不愿意提供某几道访题的信息。看起来，受访者通常用跳答或给出错误应答来处理这类冲突。例如，Moore，Stinson和Welniak（1997）在当前人口调查（Current Population Survey）中考察了收入访题丢失的数据，发现有超过1/4的工资薪酬数据丢失或不完整。这大约是日常人口登记访题数据丢失率的10倍。

有时，拒答要比低估尴尬行为频率更棘手。例如，拒答NSDUH中吸食可卡因的访题几乎等于承认自己在吸毒。简单地应答说没有吸毒可能更容易。在调查中，可能被低估的、令人窘迫的行为如吸毒、酗酒、吸烟（特别是未成年人和女性）以及流产。受访者也不愿意承认自己应该做而没有做的事儿。由此会导致过高地报告某种社会期许的行为，如投票或去教堂。

一些研究者尝试使用“宽恕式”措辞来改进对于敏感信息的报告。譬如涉及选举的访题：

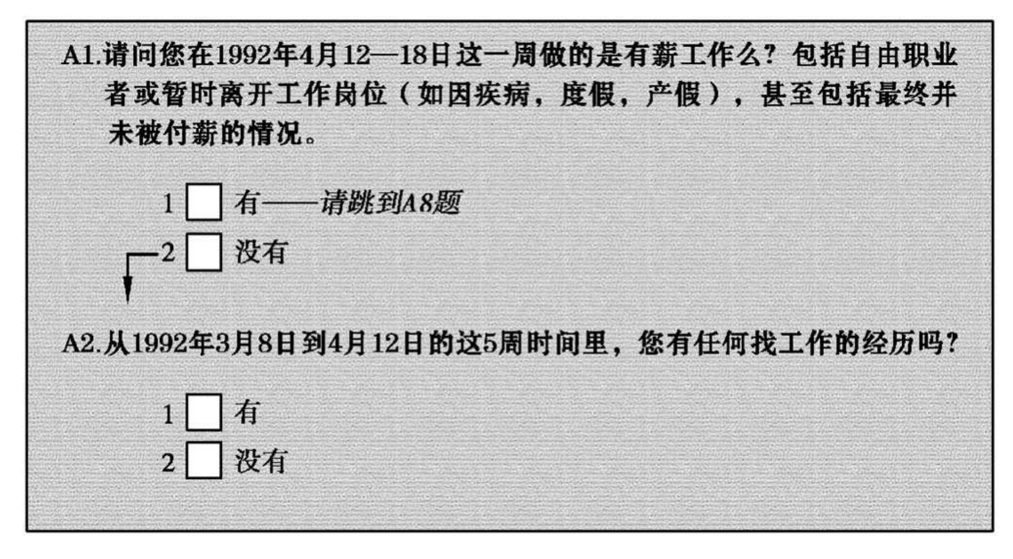
> 在和人们谈论选举时，我们经常发现许多人因为没有注册、生病或恰好没有时间而不能去参加选举。那么您呢？您在今年 11 月的选举中有没有投票？（美国国家选举研究）

这道涉及选举的访题措辞鼓励受访者报告真实的行为，即使他们没有投票。尽管如此，经验数据表明，这种措辞并没有根除对选举行为的过高报告。改进对敏感性访题应答的最有用方法可能是免去问与答过程的访员参与（参见5.3.5节 ）。一种是，可以运用纸版自访问卷，或通过计算机直接向受访者显示。相对于访员访谈，这两种形式的自访问卷都可以增加对窘迫行为的报告。另一种被称为“随机化回答技术 ”（randomized response technique）（Warner，1965），让受访者旋转转盘或使用其他随机化装置来确定他们回答敏感性还是非敏感性访题（如“您是九月份出生的么？”）。当访员不知道受访者究竟回答哪道访题时，人们似乎更乐于参与其中。在此基础上，依据随机化模型分配敏感性和非敏感性访题的已知概率来估计敏感特征

的感知性。对这项技术的实践评估可知，它的确降低了应答中的一些误差，但没有去掉所有误差。

7.3.8　引导性误差

运用自访问卷（如邮寄问卷）时，受访者必须理解访题以及问卷的任何提示语，如回答哪道访题，应答需要采取何种形式（如“勾选其中一项”），以及其他提示语。事实上，在自访问卷中，受访者的一项重要任务是弄明白答完一题后，下一题在哪里。为了帮助受访者找到完成问卷的正确路径，问卷会给出各种形式的跳答提示语（如“如果选否的话，请跳到第8题”）；在问卷中，提示语常以视觉的或图像的形式加以强调，比如黑体字和箭头。如图7.3（Jenkins and Dillman，1997）的例子。例子展示了有助于受访者完成问卷的若干原则。问卷中运用字体、图标的方式要一致。例如，访题用粗体字，与问卷其他部分相区别；题号（A1.和A2.）放在最左端的位置以吸引注意力。第一道访题（以及其后访题）的路径提示语（跳到A8题）用斜体字表示。希望由受访者填充的部分用白色，与周围较暗的背景形成了对比。另外，尽量使用箭头而不是文字来指出受访者要遵循的路径（如，对A1.题回答“否”的人被箭头引向A2.）。



A1.请问您在1992年4月12—18日这一周做的是有薪工作么？包括自由职业者或暂时离开工作岗位（如因疾病，度假，产假），甚至包括最终并未被付薪的情况。

1 □ 有——*请跳到A8题*

2 □ 没有

A2.从1992年3月8日到4月12日的这5周时间里，您有任何找工作的经历吗？

1 □ 有

2 □ 没有

图7.3　来自Jenkins和Dillman（1997）的例题

然而，受访者依然很容易产生“引导性误差”。“引导性误差 ”（navigational error）指受访者跳过应该回答的访题或回答了应该跳过的访题。受访者可能没注意到提示语，或者即使注意到提示语却没有理解。因此，相对于访员参与的调查而言，自访问卷，特别是设计较差的自访问卷，常有更高的数据丢失率。

随着自访问卷在ACASI、网络调查和电子邮件调查中的使用越来越多，需要有更多方法研究来探讨受访者对不同格式的反应。譬如问卷格式的哪种特性会减轻受访者的负担？是否不同的格式对低受教育人群的帮助要大于高受教育人群？是否精心编制格式能激发受访者完成填答？

7.4　编制好访题指南

注意到问卷设计中潜在的陷阱是有意义的。虽然用访题来说明这些陷阱更容易些，不过，建立一些避免陷阱的准则也是有益的。许多

教材曾提出过编写访题的准则。这一节将归纳Sudman和Bradburn（1982）提出的一组较为全面的准则，Converse（1986）也为编写访题提出了一些非常好的建议。事实上，Sudman和Bradburn准则并没有经受时间的考验，为此，在讨论中我们也作了一些修订。两人推荐的准则来自于经验发现，大部分是成立的，与我们已经讨论的、可能出错的环节也是一致的。

Sudman和Bradburn并没有针对所有调查访题提出一整套准则，而是针对不同类型的访题给出了编写建议：

1）有关行为的非敏感性访题。

2）有关行为的敏感性访题。

3）态度访题。

区分不同类型的访题很重要。的确，不同的访题有不同的难点。例如，敏感性访题容易产生误报，因此需要一些特殊步骤来获取准确的应答。态度访题则可能牵涉到量表，而运用量表（正如先前注意到的）有其特定的困难。我们将按顺序讨论每一类访题。

7.4.1　有关行为的非敏感性访题

对大量涉及行为的访题来说，最主要的难题是受访者可能会忘记部分或全部信息或在错误估计的基础上形成应答。因此，Sudman和Bradburn针对有关行为的非敏感性访题提出的准则都试图减轻记忆上

的负担。当然，针对非敏感性访题的准则对敏感性访题也同样适用。这些原则有：

1）在封闭式访题中，将所有可能情形都作为选项列出。

2）访题越明确越好。

3）尽量使用受访者都能明白的语言。

4）通过增加促进回忆的提示来使访题变长。

5）当有可能遗忘时，使用辅助回忆。

6）当事件经常发生且不引人注意时，让受访者保存一份日记。

7）当需要回忆较长时期发生的事件时，使用生命历程表来改进报告。

8）为降低远溯误差，要求受访者使用家庭记录或使用受约束访谈（或两者都使用）。

9）如果要兼顾成本，则可以看看过滤性访题是否有助于获得准确信息。

在这些准则中，前三种都涉及访题措辞。在应答选项中列出所有可能情形很重要，因为受访者不会主动给出没有向其明示的应答。此外，在选项的最后列出“其他”，则可能使受访者低报。例如，以下两道访题可能产生两种不同的应答分布：

1）您的种族是什么？

白人

黑人

亚裔或太平洋岛民

美洲印第安人或阿拉斯加原住民

其他种族

2）您的种族是什么？

白人

黑人

印度裔

华裔

日裔

韩裔

越南裔

菲律宾裔

其他亚裔

夏威夷原住民

关岛人

萨摩亚人

其他太平洋岛民

美洲印第安人或阿拉斯加原住民

其他种族

将“亚裔或太平洋岛民”分解为几部分，澄清了选项的含义，且使受访者更容易辨别是否应该选择。因此，在对第一道访题的应答中，可能有更多的人报告“亚裔或太平洋岛民”。

明确的访题可以降低受访者在理解上产生分歧的可能性。访题应该表述清楚其涉及的人（如“you”指的是受访者自己还是受访者家庭的所有成员）、涉及的时间段、涉及的行为等。其中常见的错误是，没有表述清楚访题涉及的参照期，如“在一周中，您一般会吃多少次甜点？”，但事实上我们的饮食习惯在一生中、一年中甚至数周内的变化非常大。

好的访题应明确参照期，如“在过去的一个月中，即从________［日期］开始，一般而言，您每周吃几次甜点？”。访员在他询问的那天会填上具体日期。

第三点建议是使用每个人都能看懂的语言。这一点说起来容易，做起来就难了。一些更详细的原则包括避免使用科技术语（如“您得过心肌梗死吗？”），而是使用日常用语（如“您有过心脏病突发的经历吗？”）；避免使用模糊的数量词（如“经常”“几乎从

不”），而是使用明确的频率选项（如“每天”“每月一次”）；不使用含糊的修饰语（如“通常”），而使用更明确的修饰语（如“大多数情况下”）。如果需要的话，也可以使用模糊的措辞或科技术语，不过，要明确定义，最好是在提问之前定义（如“心肌梗死即心脏病突发；从学术上来讲，是由心肌的某些组织死亡导致的。您得过心肌梗死么？”）。

其后的五条原则都涉及如何减少遗忘对报告准确性的影响。最根本的策略是向受访者提供更多检索提示，无论是把提示与访题相结合，还是针对某个具体范围进行提示。增加提示会使访题变长，正如第四条原则所说的，同时也会改善受访者的回忆。针对具体范畴给出提示，也就是第五条原则说的“辅助回忆 ”（aided recall）。NCVS中涉及购物的访题就是例子。访题中列出了购物的不同地点（“……药店、服装店、杂货店、五金店或便利店”），也可以对每类商店都给出一个单独访题。两种提示方法都有助于受访者唤醒记忆。让检索提示有实际作用是很重要的。如果针对子范畴提出毫无疑义的访题（如“您有多少次购买过红色的物品？”“您有多少次在下雨的周二早晨去购物？”）只会使情况变得更糟（参见Belli，Schwarz，Singer，and Talarico，2000）。还有，不符合受访者对事件编码方式的检索提示比没有提示更糟。改善回忆状况的另一种策略是让参照期的长度符合人们记忆的特性。较少发生的、对于个人情感有较大影响的、持续时间较长的事件要比经常发生的、不重要的或一闪而过的事件容易记忆。例如，相对于因感冒而去看医生，受访者更容易记住因心脏病手术而住院。因此，容易记忆的事件如住院，参照期应该较长（一年可能较为合适）；而不容易记忆的事件如看医生，其参照期应该较短（两周或一个月也许较为合适）。

改变参照期的方法也有局限性。当调查涉及日常的或没什么影响的事件时（如小宗消费购物或食物摄入），参照期应该很短，如回忆昨天发生的事情。在这种情况下，与其依靠受访者的回忆，不如要求受访者保存一份日志。另一方面，调查不得不运用较长的参照期。例如，追踪调查不可能太频繁地去访问受访者，通常是几个月或一年拜访受访者一次，因为没有那么多经费。因此，必须使用较长的参照期以覆盖在两次访谈之间的时间间隔（如NCVS每隔六个月去拜访一次样本人群的住所）。当必须使用较长参照期时，用生命历程表可以改善回忆状况。生命历程表记录了人一生中若干时期发生的重大事件，如结婚生子、工作和住房等。在历程表上的这些记录有助于唤起受访者对更普通事件的记忆，如那个时期的收入、疾病、犯罪等。生命历程表上的自传体式重大事件为检索提供了丰富的时间提示和主题提示（Belli，1998）；生命历程表还可以作为时间标记来帮助回忆其他事件发生的具体时间。生命历程表要求访员与受访者之间进行不那么结构化的访谈。无疑，这些工具增加了访员效应导致的调查结果变异性（参见9.3节 ）。这是调查方法研究中较为成熟的一个领域。

另一种记忆误差是误记了事件发生的时间（即“远溯误差”）。第八条原则提供了降低此类误差的两种方法。第一种方法是让受访者通过查询家庭记录（日历、支票簿、票据、保险单或其他财政记录）来帮助回忆购物、看医生或其他任何可能留下文字材料的事件，并确认事件发生的具体日期。第二种方法是“约束”，即向受访者提醒前次调查（即在追踪调查的前一轮调查）中已经报告过的事件。

最后一条建议是，在数据搜集成本较高的情况下，考虑使用代答者提供信息。“代答者 ”（proxy）指除了原定受访者之外的任何人。大多数调查会要求父母提供孩子的信息，而不会直接访问孩子。

其他调查也许会让家庭的某个成年人来报告家庭其他人的情况。允许代答，能降低成本，因为访员可以当场搜集信息，而不用安排回访。同时，代答与受访者应答之间会有系统差别。例如，相对于受访者应答依靠偶发事件（如某事的详细记忆），代答者应答会更依赖于一般事件信息（即通常发生情况的信息）。另外，代答者和受访者可能对某事知道的也不一样。让父母代替他们的孩子应答是否吸烟是没有意义的，孩子们通常会隐瞒吸烟行为，特别是对父母。不过，总体而言，代答者通常也能提供可靠的信息（参见O'Muircheartaigh，1991）。

7.4.2　有关行为的敏感性访题

正如在7.3.7节讨论的那样，有些调查中有一些涉及非法或令人尴尬行为的访题，譬如吸食可卡因、酗酒、吸烟等。NSDUH和BRFSS中有许多这样的访题。下面是Sudman和Bradburn归纳的编写行为敏感性访题的原则：

1）在问及敏感行为发生频率时，使用开放式而不是封闭式的结构。

2）使用长一些的访题，访题不要过于简洁。

3）使用人们熟悉的语言来描述敏感行为。

4）精心导入访题，以避免误报。

5）先询问较长时期（如某人的一生）或较远的过去发生的敏感行为。

6）把某个敏感访题放到其他敏感问题中，使其变得不那么显眼。

7）使用自访问卷或类似方法来改进报告。

8）考虑运用日志搜集信息。

9）在问卷结尾设计若干题目来评价关键访题的敏感程度。

10）搜集有效的数据。

相对于封闭式访题来说，涉及敏感行为的开放式访题有两个优点：首先，封闭式访题不可避免地丢失信息（对较为极端的人群而言）。其次，受访者可能认为封闭式访题的选项是总体中行为分布的某种暗示，进而影响受访者的应答（参见文本框中Schwarz及其同事的讨论）。

Sudman和Bradburn（1982）建议使用长一些的访题，是为了引发更为完整的回忆（通过给予受访者更多回忆的时间）。这种方法对有被低估倾向的行为尤其有效（如酗酒）。例如，如果询问受访者饮酒量，建议使用如下的措辞：

> 最近几年，酒变得越来越流行了；说到“酒”，指的是利口酒、浸果酒、雪利酒及类似的酒精饮料，还有像淡酒、汽酒和香槟等。您曾经喝过酒或香槟么？哪怕就一次。

对于各类酒的罗列，有助于解释研究范畴的界限，更多的是提供检索的额外时间。下一条建议，使用人们熟悉的语言来描述敏感行为（如“性交”而不是“交媾”），不仅会让受访者看访题时更舒服一些，而且有改善回忆的倾向，因为访题中的措辞与人们对事件编码的措辞一致。在访谈开始时，访员可以确认受访者喜欢使用哪些词。

“导入 ”（loading）访题指的是，以引出某种特定应答方式（这里指社会并不期许的行为）来编写访题。Sudman和Bradburn（1982）讨论了几种策略。“众行”法，如“即便是最为镇静的父母，有时也会对孩子大发雷霆。在过去的一周，您的孩子有没有做过让您觉得生气的事情？”；“假定”法，如“在过去的一周，有多少次您孩子做了让您觉得生气的事情？”；“诉诸权威”法，如“心理学家认为父母们宣泄自己的抑郁很重要。在过去一周，您的孩子有没有做过让您觉得生气的事情？”；以及“有理”法，如“父母有时会因为疲倦、心烦或孩子总是淘气而生气。在过去的一周，您的孩子有没有做过让您觉得生气的事情？”。前面讨论的选举访题就运用了最后一种方法。

后面的两种建议有助于降低访题的敏感性。一般来说，承认曾经做过或在很久以前做过某事总比承认在最近做过不那么让人尴尬。例如在2000年美国总统竞选中，候选人小布什承认他在十多年前有过酗酒行为，却并没有引起公众多大的反应。但如果他承认自己在前一天还有酗酒行为，那就是另一回事了。大多数研究者认为，敏感性访题不应该出现在访谈的开始，而应该出现在一些不太敏感的访题之后。另外，将一个敏感性访题（例如从商店偷东西的访题）放到其他更为敏感的访题中（如武装抢劫），有助于通过比较来降低访题的敏感度。正如许多判断一样，人们感受到的敏感性会受到背景影响。

正如已经注意到的那样，改善敏感行为报告的有效方法之一是让受访者自访或通过计算机填写应答。另一种方法是运用随机化方法让访员不知道受访者在回答哪道访题。譬如，受访者从盒子中取出红色的珠子或蓝色的珠子，由此来确定受访者需要应答的访题：

（红色）在过去12个月中，您曾因酒后驾车而被捕过么？

（蓝色）您的生日是在六月么？

访员记录“是”或“否”的应答，而不知道受访者回答的是哪道访题。由于研究者知道红珠或蓝珠被选择的概率（以及受访者在六月出生的概率），因此可以估计对酒后驾车选“是”的比例。

第三种方法，即让受访者保存一份日志。这种方法在自访问卷中可以有效减轻记忆上的负担。日志法常常需要详细记录，由此也会导致较低的受访效率。

最后的两个建议是通过评估受访者应答时的不安程度来评价访题敏感程度，并且通过与外部数据进行比较来评价应答的整体准确性。如把受访者对最近接触毒品的应答与尿检结果比较。

7.4.3　态度访题

许多调查会问及受访者的态度。在我们的6个调查范例中，只有SOC包括了大量关于态度的题目。然而，关于态度的问题是很常见的一

类问题，Sudman和Bradburn也专门为它们提出了一些原则。我们在下边列出这些建议的一个修正版：

1）明确态度研究的目标。

2）避免双重目的访题。

3）测量态度的强度，如必要，可专门设计访题。

4）在不丢失关键信息的情况下，尽量使用两极访题。

5）认真考虑题干的措辞，因为措辞对受访者应答有重大影响。

6）为测量态度随时间的变化，每次都询问相同的访题。

7）当需要同时询问某个主题的一般和特殊信息时，首先问一般信息。

8）当需要问及有多种情形时，先问人们较少考虑的情形。

9）使用封闭式访题来测量态度。

10）使用5级或7级量表，并对每一个选项进行标注。

11）从不易被选择的一端开始。

12）使用模拟方法（如使用“情感温度计”）来收集更详细的信息。

13）只有当受访者能看到所有选择时，才可使用排列方式；否则，使用配对比较。

14）对每个主题都作出评价；不要使用多项选择题。

前面的六条原则涉及措辞。第一条，明确态度研究的目标。看看下面这道题：

您认为政府在反恐措施上花费的精力过少、适度，还是过分？

目标明确可以改善受访者对访题的理解，使受访者对反恐措施以及政府精力花费的解释更为一致。双重目标（double barreled items）的访题希望同时考察对两种事物的态度，这是要避免的。例如，下面的第一道访题将对流产的态度和对最高法院的态度弄在了一起，而第二道访题则将对流产的态度和对女性权利的态度搅在了一起：

1）美国最高法院裁定女性有权在怀孕的前三个月终止怀孕。您支持还是反对这个裁定？

2）您是否因为合法流产给予了女性选择权而赞成它？

对于双重目标访题的应答，很难给予解释，即到底是对哪个事物的态度？

对态度，研究者关心的特征有两个，即方向（支持或反对，正面或反面，赞成或不赞成）和强度。第三条建议涉及的是态度的强度，通过运用量表（“非常反对”“有些反对”等）获得对强度的测量。单个访题或一组访题，都可以测量态度的强度。第四条建议是在不丢失关键信息的情况下尽量使用两极测量。如在同一道访题中询问相互冲突的政治诉求，而不是分拆为两道访题：

> 是政府应该负责每个人有适当的医疗保障还是个人自己应该负责?

“两极法 ”（bipolar approach）强令受访者在可能的选择之间作出选择，从而杜绝了访员不表态的情况。不过，这种方法有时会漏掉某些微妙的信息。例如，积极或消极的感情并不总是强（负）相关的，因此，针对让受访者高兴或沮丧的事物，设计单独的访题是有意义的。第五条建议与纳入中间选项（如“既不赞成也不反对”）或无意见选项的后果有关。一般来说，应该把这类选项纳入访题，除非有非常正当的理由不纳入（如在选举民意测验中，通过让受访者倾向一方或另一方来表达选择偏好是很重要的；这种情况下，可以省去中间选项或无意见选项）。下一条建议表明，测量态度变化的唯一方法是进行相同事物的比较，即在不同时点使用相同的访题。

接下来的两条建议是希望减轻访题顺序的影响。如果问卷同时包含了对同一主题的一般访题和特殊访题的话，最好是先询问一般访题；否则，对一般访题的应答可能受前面特殊访题的数值或内容的影响。（回忆一下前边的讨论，如果在婚姻幸福程度之后提问总体幸福程度，则受访者将重构整体幸福程度的应答。）如果问卷中有若干道

人们在熟悉程度上有差异的访题（例如，GSS中包含有数道在不同情境下询问人们对流产支持程度的访题），那么，跟着的生疏访题可能会让受访者觉得更生疏了。在这种情况下，把生疏的访题前置，可能会得到更准确的应答。

最后六项建议涉及态度访题中几乎无处不在的量表。第一条（即建议中的第九条）建议使用封闭式访题，而不是开放式访题，因为开放式访题太难进行编码。下一条建议就更有针对性了，建议使用5级或7级量表。过少分级的量表会丢失信息，而过多分级的量表又会产生认知负担。语言标注有助于确保受访者以相同方式来理解量表。当受访者对选项的熟悉程度不同时，生疏选项只有在第一个出现时，多数受访者才会考虑。如果访员向受访者大声读出访题，则受访者又可能首先考虑他们听到的最后一个选项；在这种情况下，生疏选项应该放到最后。

还有一些常见的、其他的态度访题形式。最后三项建议涉及对类比（如运用体温计）、排序以及多项选择的运用。当要运用7级以上的量表时，**类比法** （analogue method）可能有助于减轻受访者的认知负担。例如，“情感温度计”要求受访者用0（表示非常冷淡）到100（表示非常喜爱）来评价他们对于公众人物喜爱的程度。研究结果表明，有13个左右的数字是受访者较喜欢选择的。受访者不习惯做不确定的数字判断（这也是受访者倾向于选择以0或5为结尾数字的原因），因此最好在量表中包括上限和下限，正如运用情感温度计那样。有时，还会要求受访者对不同事物排序 （ranking）（如希望孩子的品格）。正如第十三条建议指出，除非把需要排序的所有事物都列在同一张给受访者查看的卡片上，否则，就会超出受访者的认知能力。如果不能给受访者查看，譬如电话访谈，研究者就得采用两两比

较的方式。最后一条建议是不鼓励研究者使用多选访题（check-all-that-apply），因为受访者只会选择与其相符的选项（Rasinski，Mingay，and Bradburn，1994）。拆分内容，让受访者对每个主题都作出是或否（赞成或不赞成，支持或反对等）的评价，就能减少这类误差。

7.4.4 自访访题

Sudman和Bradburn（以及其他大多数问卷设计教材）关注的是面访和电话访问的问卷。在面访和电访中，接受过训练的访员充当了受访者和问卷之间的媒介。相比之下，邮寄问卷调查则要求受访者自己理解访题、提示语，以及理解应答方式。Jenkins和Dillman（1997）（同时参见Redline and Dillman，2002）提出了若干建议，以提高受访者正确填答邮寄问卷和其他自访问卷的准确率。以下是他们提供的建议：

1）使用统一的视觉要素标注完成问卷的路径。

2）当必须改变惯例时，用醒目的图标提醒受访者注意。

3）运用指导语，并把指导语放在该放的位置。

4）将需要配合使用的信息放在同一位置。

5）一次只问一个问题。

第一条建议说的“视觉要素”包括亮度、颜色、形状以及在页面的位置。正如在7.3.8节提到的那样，在问卷中，可以将访题的题号放在页面的最左端，用黑体字标示访题的题号和文字，用与访题不同的字体来显示提示语，或使用图标（如箭头）指引应答的接续和跳转。如果问卷自始至终都使用同一套图示，则受访者就会很熟悉。不幸的是，有时候很难做到自始至终。例如，访题的应答形式可能会发生改变。在这种情况下，根据第二条建议，提示必须非常显眼，使受访者做该做的事。图7.4展示了从圈答到填答的转化，空白应答区和阴影背景之间的对比，让受访者注意填答方式的改变，做该做的事。

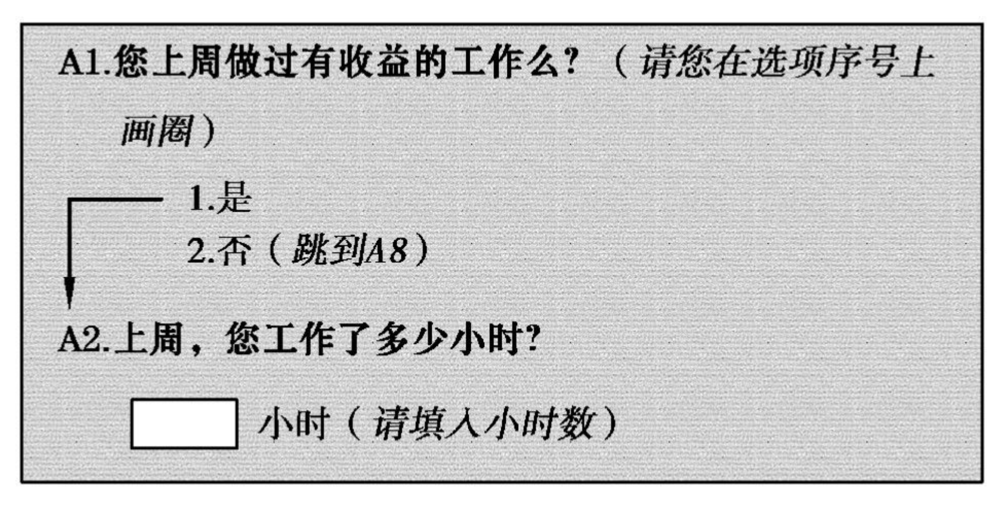


图7.4　高亮应答框产生的视觉对比

下面的两条建议与提示语的位置有关。自访问卷的第一页通常会对问卷的应答方式进行说明。Jenkins和Dillman认为，受访者可能会不看第一页的说明就开始应答，即使读了说明，等到真要用的时候也可能忘记了。受访者更可能注意到的是及时运用的提示语。在图7.4中，应答提示语（请您在选项序号画圈）就列在了选项之前。相关的建议是，把概念相近的信息排列在一起。例如，让受访者不必先看访题再看标题才知道如何应答。访题应答需要的全部信息都应该放在一处。

调查访题常试图用一道访题覆盖多个可能性。例如，如前涉及问医的访题就比较复杂，在一道访题内，试图把医生、其他医护人员、当面咨询及电话咨询都包括了进来。在自访问卷中，用一道访题问多个主题的诱惑性很强，有时增加一两道题就意味着增加一页纸。同时，询问多个主题又会造成受访者理解的负担，甚至因此不能保持清晰、完整的思路。下面是Jenkins和Dillman讨论过的一个例子：

> 在您的雇员中，有多少是全职员工且有健康保险福利？有多少是全职员工却没有健康保险福利？另外，各类型雇员为公司服务的平均时间是多长？

这道访题实际上询问了四个单独的问题：有多少全职员工有健康保险福利？平均来说，他们为公司服务的时间有多久？有多少全职员工没有健康保险福利？平均来说，他们为公司服务的时间有多久？不管这类访题节省了多少问卷空间，可由此带来的益处则可能被理解上的损失所抵消。

7.5 小结

受访者在应答过程中，涉及理解、记忆检索、判断和估计，以及报告等一系列的行为。有些通过视觉呈现给受访者的自访问卷还需要受访者自己对完成问卷的流程作决策。对行为和态度的测量，甚至在判断和估计阶段就有特殊的麻烦。

在调查方法的文献中，有一些涉及随机化实验。这些实验用来说明应答过程各个阶段的测量误差，如编码错误，即信息不以可读取的

形式储存在记忆中；因语法或措辞而导致误解；遗忘以及其他记忆上的困难；在涉及行为发生频率时的低估；以及由访题上下文和敏感程度而产生的判断效应。随时间的推移，调查方法研究已经发展了处理这些难题的工具，且因是否为非敏感性行为或态度而不同。

细心的读者可能会发现，本章用了近两倍于其他讨论的篇幅来讨论影响应答的因素及其解决方案，最主要的原因是每一条解决方案的建议都有局限性。因此，我们认为讨论这些建议产生的基础比建议本身更重要。

这些建议的局限是什么？首先，对于任何一组建议而言，不论概括得多么全面，都不可能适用于所有情形。例如，有些研究者认为某类访题很难产生可靠应答，如因果关系访题和对假设情境反应的访题。我们列出的建议就忽略了这一讨论。因此，这些建议不可能覆盖每种情况。其次，每条建议适用的情景都有例外。Sudman和Bradburn建议不要使用多选题，我们也大体同意这项建议，不过，在2000年人口普查的简表中，还是使用了多选题来搜集种族信息。大多数研究者认为，2000年人口普查中的方法比询问是白人（是或否）、黑人（是或否）或类似的选项要好。多选题给予受访者自然和有效的方式来表明其多种族背景。这些建议的另一个局限是，有时会相互冲突。一方面要求访题明确定义态度研究的目标；另一方面又要尽量避免使用双重目标访题。我们揣测，正是为了明确态度研究的目标，才产生了许多有双重目标的访题。同样，把模糊的日常用语说清楚（如“看医生”），就得使用较多的复杂句（参见前面对问医的讨论）。这些建议之间的相互冲突是个严重问题。这些冲突也代表着在调查设计中对同等重要的各因素之间的平衡。例如，在态度访题中包含中间选项有助于让确实处于中间状态的受访者准确表达自己的观点；缺陷是为那

些不想认真应答的受访者提供了一条捷径。很难说这两种不同的考量哪个更重要。另一种情况是，把复杂的主题拆分为若干小访题也许能改进应答，却增加了问卷的长度。

总之，这些建议只是在特定情形下才更为有效。在应用时，应尽量检验是否符合实际情况。即便是最有经验的问卷设计者，也需要数据来帮助他们去设计问卷。毕竟，实践是检验真理的唯一标准。在下一章，我们会讨论检验问卷的方法。题，请确认如下事宜：①来自于哪些调查的哪些访题；②如果要修改，则说明修改方法；③怎样通过修改以满足研究目的。注意：已经使用过的或为公众了解的访题不一定就是好访题。在问卷编制中，应运用本章提供的经验和建议。

关键词

编码（encoding）

检索（retrieval）

估计（estimation）

报告（reporting）

社会期许（social desirability）

语法不明确（grammatical ambiguity）

错误预设（faulty presupposition）

模糊量词（vague quantifiers）

错误推断（false inference）

复述（rehearsal）

一般记忆（generic memory）

参照期（reference period）

回忆并计算（recall-and-count）

基于印象的估计（impression-based estimation）

正面偏差（positive bias）

近呈效应（recency effect）

随机化应答技术（randomized response technique）

辅助回忆（aided recall）

导入（loading）

两极法（bipolar approach）

排序法（ranking）

理解（comprehension）

检索线索（retrieval cue）

判断（judgment）

默许（acquiescence）

满足（satisficing）

过分复杂（excessive complexity）

模糊概念（vague concepts）

陌生措辞（unfamiliar term）

综合社会调查（general social survey）

检索失败（retrieval failure）

重构（reconstruction）

远溯（telescoping）

基于频率的估计（ratebased estimation）

开放和封闭访题（open and closed questions）

首呈效应（primacy effect）

敏感性访题（sensitive question）

引导性误差（navigational error）

代答者（proxy）

双重目标（double barreled items）

类比法（analogue method）

多选题（check-all-that-apply）

进一步阅读资料

Sudman, S., and Bradburn, N. (1982), *Asking Questions : A Practical Guide To Questionnaire Design* , San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N., and Schwarz, N. (1996), *Thinking about Answers : The Application of Cognitive Processes to Survey Methodology* , San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000), *The Psychology of Survey Responses* , Cambridge: Cambridge University Press.

作业

1. 为测量公众对能源储备的支持程度，（虚拟的）安德鲁斯基金会（Andrews Foundation）进行了一项民意测验，发现72%的美国人支持如下的表述（访题要求给出“赞成”或“不赞成”应答）：

   我支持奥巴马总统使用美国军队以帮助地方城市安装更多高能效公共照明设施来实现能源自立。

   （a）访题测量的是态度的哪个重要方面？

   （b）满足用于评估通过安装更多高能效照明设施来实现能源自立的研究目标吗？解释满足或不满足的原因。

2. 用你掌握的编制态度访题的知识，写一份标准的、由访员实施的问卷，测量公众对出动地面部队侵略伊拉克支持的方向与程度。可以使用多道访题以及任何需要的选项形式。请务必明确选项的类型（如果使用选项的话），以及需要向受访者读出的内容（与对访员的提示语相对应）。用适当的格式或标示，指出任何需要跳过的访题。

   可以采用多种方法来完成这项任务。开始时，可以先搜索在实际调查中使用过的类似访题（例如，看一下盖洛普［Gallup］或皮尤［Pew Research Centre］的网站）。如果要借用已有的访

3. 写1～2段文字说明如何在编制问卷中运用了本章提供的建议，尽可能详细。例如，如果采用多道访题，则可以说明为什么选择某道问题而不是其他作为第一道访题？如果运用了11级量表，则说明理由。还有，在多种选择面前，你是如何取舍的？这些取舍会如何影响应答结果？

4. 社会期许如何影响应答？当调查宗教活动参与时，请列出两种降低社会期许影响的方法。

5. 想一想应答中的认知处理模型，列出下题题干措辞及选项中存在的问题：

   许多有车的人会对自己的汽车定期做养护，如换油等。通常您如何进行汽车养护？指出其中的一项或两项。【向受访者出示选项，并作选择】

   换油　　　　　　换液　　　　　　调整发动机

车体维修　　　　保修服务　　　　轮胎保养

传动装置的维修　　　　空调维修

6. 指出下列访题中存在的问题并提出修改方法。

练习1

在过去4周中，即从【四周前的日期】到今天，您做过家务吗？如清洁、烹饪、园艺、房屋修理，不包括工作。

练习2

在过去1周中，您饮用酒精饮料多少次？

练习3

在您居住的区域，如果要满足必要的花费，您和您家庭每月需要的最低收入（不包括任何扣除）是多少才能收支相抵？

练习4

在过去12个月中，即从【日期】开始，您大概有多少天因病伤而卧床半天以上？包括您在医院住院的情况。

7. 简要地给出避免使用“同意—不同意”形式的两点理由。

8. 指出在下列情形下，受访者会采用何种评估策略，以及带来何种频率报告。

（a）在过去2年中，受访者住院的次数。

（b）在过去1个月中，受访者在餐馆就餐的次数。

（c）在上个夏天，受访者的配偶/同伴度假的次数。

9. 简述研究者处理远溯的两个策略。

8　调查访题评估

8.1　导言

第2章讨论了调查统计误差产生的两类主要来源：非观察误差（如果受访者特征与总体特征不匹配）和观察误差或测量误差（如果对访题的应答不能满足调查目标）。本章讨论调查访题的评估方法，并考察有多少测量误差是由访题带来的。

访题评估包括两个方面。第一，评估第7章讨论过的如访题有多好理解、应答有多困难、对访题的理解和应答质量又如何影响了测量质量等问题。调查研究者通过观察受访者对访题的理解和应答来评估受访者对访题的理解力、记忆检索困难以及相关的议题。背后的假设是，相对于不容易懂或因为其他原因难以应答的访题，如果访题通俗易懂，受访者就不会产生认知困难，带来的测量误差也更小。第二，访题评估就是评价应答与测量之间的匹配程度，即直接估计测量误差。基本方法是，要么与其他测量进行比较，要么进行重测。与其他测量进行比较，检验的是效度或应答偏差；重复测量，检验的则是信度或应答方差。在第8.9节，我们会讨论这些误差估计的操作技巧。

这里有三条重要标准是所有调查都必须满足的：

1）内容标准（content standards）。（如访题是否问了要问的内容？）

2）认知标准（cognitive standards）。（如受访者是否能明白无误地理解访题？受访者是否按照要求应答？受访者是否愿意并能形成应答？）

3）可用标准（usability standards）。（如受访者或访员［如果有的话］是否很容易完成应答或倾向于完成应答？）

与这三个标准相关的是不同的评估方法。在讨论访题评估的主要方法之后，我们将回到每一种方法，来看其在三个标准的意义上提供什么信息。

在过去的20年里，调查方法研究的一项主要进展就是日益注重对访题的系统评估。用于评估访题草案的方法，至少有以下五种：

1）专家评估。主题专家（subject matter experts）评估访题内容是否适合用来测量要测的概念，或问卷设计专家评估访题是否满足上述三条要求。

2）焦点小组。调查设计人员与目标人群进行半结构式的（“专题”，focused）讨论，探求问卷应该覆盖哪些议题，如何思考这些议题，以及用什么措辞来表达这些议题。

3）认知访谈。访员用访题草稿访问受访者，观察受访者如何理解访题，以及如何形成应答。

4）试调查。用与正式调查相同的方法获取样本，让访员对少量样本（不超过100人）进行调查，在这个过程中要：（a）给访员介绍整个调查（让访员了解提问与应答的具体情形）；

（b）用数据表呈现数据，用以观察可能存在的问题（如缺失值比率比较高的选项）；（c）记录访谈和行为编码（为一些不易措辞的表达提供数据，参见Fowler and Cannell，1996）。

5）随机或折半实验。用不同比例的样本测试针对同一测量访题的不同措辞（Fowler，2004；Schuman and Pressser，1981；Tourangeau，2004）。

下面各节分别就如何在实践中实施评估进行讨论。

8.2 专家评估

正如前面提到的，主题专家和问卷设计专家都可以对问卷草稿进行评议。这里我们专注于问卷设计专家的作用，同时强调，由相关专家进行问卷评估的目的是确保用问卷能够搜集到调查分析目标不可或缺的内容。专家评估的内容包括问卷措辞、结构、应答选项、访题顺序、指导语和访员引导，以及问卷应答流程。

有时，专家会运用常见问题列表。从Payne（1952）开始，多年以来，不少作者发表了编制好访题的原则。最初，这些原则基本上是作者的观点和判断。随着时间的发展，认知测试、试调查行为编码以及心理量表逐渐加了进来。第7.4节曾经给出过我们自己的原则（引用Sundman and Bradburn，1982）。

这些原则面对的共同问题是，访题易受解释和判断的影响。因此，判断访题好坏的一个基本标准是受访者的理解一致，且理解的意

思与设计者的用意一致。尽管对原则没有异议，可在面对具体访题时却总有不同的意见。因此，运用问题列表进行对照检查是评估访题的基本方法，同时，问题列表也可以用于其他测试。

人们进行了很多努力，试图制作一套细致的、具体的、用于评估访题潜在问题的列表。Lessler和Forsyth（1996）通过对应答过程的认知分析，提出了一个25类问题的列表，与第7.2节提出的异曲同工。Grasser，Bommareddy，Swamer和Golding（1996）区分了12类主要问题，他们还开发了把这些问题用于访题评估并自动给出评估结果的计算机程序（Grasser，Kennedey，Wiemer-Hastings，and Ottati，1999）。Grasser，Bommareddy，Swamerand Golding提出的12种问题包括：

1）句法复杂。

2）记忆内容超负荷。

3）名词短语模糊或模棱两可。

4）专业词汇生僻。

5）陈述或相关措辞模糊或不精确。

6）假定有误或错误。

7）问题类别不清。

8）多种类别混合。

9）访题分类与应答选项不匹配。

10）难以应答（如回忆困难）。

11）受访者无从应答。

12）访题目的不明确。

Grasser等人提出这些问题的假设是，受访者会首先判断访题类型（如“是什么”还是“为什么”）。因此，如果访题类型不清或至少有两类，就给受访者制造了麻烦。

8.3　焦点小组

在开发调查工具之前，研究者经常会招募几组志愿者参与调查主题的系统讨论。“焦点小组 ”（focus group）是由6～10人参与的、有人主持的讨论（参见Krueger and Casey，2000）。在焦点小组讨论中，鼓励成员表达自己的观点，即使他人有不同的意见，也不会觉得不舒服，小组成员之间的观点也可以相互影响。

研究者尽心竭力来使讨论的话题结构化，试图发现讨论主题的关键议题。在开发调查工具的早期，焦点小组可以帮助研究者了解目标群体如何理解问卷中的概念，即在讨论中使用哪些术语，在一些关键议题上，目标群体选取了概念的哪些维度，等等。焦点小组还可讨论对不同抽样方法的反馈以及调查资助方的意见。

焦点小组的主持者需要尽力创造一个开放、轻松、自由的氛围。好的主持者能巧妙地让小组讨论有的放矢且在不同话题转换时过渡自然；好的主持者要鼓励所有成员发言，让腼腆者战胜羞怯，请喜欢发

言的人适可而止；好的主持者还要全神贯注地倾听每位参与者的讨论，及时捕捉其他成员的反应；好的主持者还需要依据讨论主题因势利导。因此，焦点小组的主持者必须明确研究目标，对讨论中提出的问题给予明确的反馈，而不是照本宣科。

通常，选择焦点小组成员时会找针对某个话题关键维度上有相似或关联的人。例如，在讨论就业议题时，就需要把在意寻找长期职位的与寻找零工的分开。如果调查覆盖不同子群体，则分子群体组织焦点小组会获得各自的观点与表述。

有时，也需要在一个特殊的房间举行焦点小组讨论。在这个房间里，有一个单面镜，研究小组可以在单面镜的后面观察讨论；也有音频、视频记录设备，以记录焦点小组的讨论。焦点小组讨论的结果有时是关键点笔记，有时则是整个讨论的笔录。有视频记录的，也可以对要点进行编辑汇总。

焦点小组是问卷开发早期的常用工具。通过焦点小组，可以了解到受访者对调查主题的知晓情况。如此，焦点小组对问卷设计者的吸引力主要有三个方面：

1）焦点小组是了解受访者对调查主题知晓情况以及如何组织知识的有效方法。例如，在健康保险调查中，了解受访者知道有哪些类型的保险以及每一类保险的内容（或自认为了解什么内容）就非常有用。另外，了解受访者关心什么（或不关心什么）议题或维度也非常有用。最后，了解受访者如何思考研究主题以及对研究主题如何分类或分组也非常有益。例如，在健康保险调查中，受访者可能认为健康维护组织

（HMOs）与其他健康服务计划大不相同。这类信息可以帮助研究者对问卷进行结构化处理，以提供最为准确的报告。

2）焦点小组访谈也是确定受访者措辞以及如何理解研究主题的好方法。在设计访题时，一个主要目标就是尽量使用受访者熟悉并有一致理解的术语。焦点小组为识别应答者讨论研究主题的术语以及他们对候选词汇和术语的理解提供了绝佳机会。

3）调查访题询问的是受访者了解的事情。无论是询问主观状况（诸如感觉、观点或感受），还是客观的环境或经历，如果研究者对受访者要报告的内容有充分的把握，即抓住了真实的状况，那么，访题的设计就趋于完美了。因此，焦点小组的最后一个用处就是向研究者传递受访者在谈到研究主题时一定会表达的信息。在焦点小组讨论中，通过话题引导，把一项调查要覆盖的话题做充分讨论、感受受访者的经验以及可能给予的应答，可以让研究者编制出适宜调查环境的访题。

焦点小组的独到之处在于卓有成效地获得一组人的反馈。尽管如此，焦点小组也有三点不足：

1）焦点小组讨论的参与者可能不能代表受访者群体，因此，不能把从焦点小组获得的感受和经验进行简单外推。

2）焦点小组并非对具体访题措辞进行评估或发现受访者应答形成过程的场合。焦点小组能让研究者感受到不同的受访者类型

和范围。然而，如果要评估访题的措辞以及受访者对访题的认知，更简单的方法是采用一对一的测试。

3）从焦点小组获得的信息很少有定量的，如此，焦点小组结论的信度不高，也很难重复，且易受焦点小组人员判断的影响。

尽管如此，在编制问卷之前，焦点小组还是从目标群体搜集涉及调查主题信息的有效途径。

8.4 认知访谈

美国国家研究理事会（NRC）1983年把认知心理学家和调查设计方法专家找到一起开了一个研讨会，希望发现二者潜在的共同兴趣。研讨会的后果之一就是，调查研究者开始挖掘认知心理学家的一些技术的价值，以探讨受访者如何理解与应答（Jabine，Straf，Tanur and Tourangeau， 1984；Schwarz and Sudman， 1987）。Schuman和Presser（1981）以及Belson（1981）提供了前些年调查访题被相当程度误解的例子；NRC研讨会产生的论文更激发了研究者探讨受访者如何理解和应答的兴趣。

那次研讨会上讨论的一种方法就是，使用认知访谈来测试访题。*认知访谈*（cognitive interviewing）的基础是Simon及其合作者（参见如Ericsson and Simon，1980；1984）发明的“*方案分析*”（protocol analysis）技术。“方案分析”通过让受访者把自己在受访中的思考过程说出来，并由研究人员进行记录。Simon对人们解决不

同问题的过程很感兴趣，如证明简单的数学定理或下棋。“认知访谈”覆盖了整个认知过程，包括：

1）想什么说什么（受访者在应答时，把自己的所想说出来）。

2）回顾所想与所说（在提供应答后或访问结束后，让受访者描述获得应答的过程）。

3）信心评级（让受访者对自己应答的信心评级）。

4）释义（让受访者用自己的话陈述访题）。

5）定义（让受访者对访题中的关键术语给出定义）。

6）追问（让受访者对后续访题应答，以揭示受访者的应答模式）。

这是从Jobe和Mingay（1989）的列表中选出的；亦可参考Forsyth和Lessler（1992）的工作。

正如列出内容所示，认知访谈不是一个简单方法。通常，受访者都是招募的有酬志愿者，研究者则用访题草稿以及后续追问访题，试图揭示受访者理解访题、提供应答的过程。在访问中，或许会要求受访者把自己应答过程的所思所想说出来。访谈者可以是研究者、认知心理学家、调查方法学专家、接受过特殊培训的访员或一般访员。

论试调查方法的替代

Presser和Blaire（1994）比较了四种试调查方法。

研究设计：在草拟和修订中，试调查人员独立对普通“测试”访题的140个选项进行评估。第一，采用传统试调查方法，让8位电话访员在第一轮做35份调查，在第二轮做43份调查。第二，研究者对访问的行为编码进行研究。第三，让3位认知访员运用“随后追问”和“想什么说什么”技术对30位受访者进行访谈。第四，问卷专家对问卷进行两轮问题识别。

研究发现：在常规试调查、认知访谈，以及行为编码中识别出了90个问题，而专家讨论则识别出了140个问题。试调查和认知访谈表现了很大的变异性。行为编码和专家讨论对问题的识别则相当稳定。试调查和行为编码长于发现有访员调查中的问题，认知访谈则长于发现理解中的问题和无访员调查中的问题。相对而言，专家讨论则是最有效的方法。

研究局限：在评估中仅运用了一套问卷。电访试调查的结果不一定可运用于面访。对问题的重要性没有区分。还有，没有考察研究者解决问题的能力。

研究意义：研究结果说明，在问卷开发中应多用专家讨论方法。

在认知测试中，不同组织和不同访员采用的技巧不同，或综合不同技巧来搜集信息。有人用成熟的方法，另一个人可能强调“想什么说什么”。有人会在受访者提供一道访题的应答之后马上要求受访者进行回顾，有人则会等到访问结束后再请受访者回顾。同样，对认知

访谈的记录方式也不相同，有的正式（如对访谈进行录音、录像并整理记录），有的则非正式（如仅在访谈期间记笔记）。

尽管认知访谈的运用日益增多，访谈技巧也渐趋有效，更为迫切的则是把从认知访谈中获得可信的发现、改善数据质量的价值以及变异显著性用于经验研究。DeMaio和Landreth（2004）进行了迄今为止最全面的研究，结果表明：用不同方法进行认知测试的结果大致相同。尽管如此，他们的研究还是表明，在评估相同访题时，三个小组在识别访题的问题、认定问题、解决问题等方面依然存在相当大的差异。另外，观察研究表明，不同访员对认知访谈中发生情况识别也有很大差异（Beatty，2004）。另一方面，Fowler（2004）的研究表明，用认知访谈后修订的访题所获得的数据，质量有明显改善。只是，认知测试在多大程度上能整体性改善数据质量，证据还非常有限（Willis，Demaio，and Harris-Kojetin，1999；Forsyth，Rothgeb and Willis，2004）。

8.5　实地测试和行为编码

“试调查 ”（pretest）是在开展正式调查前进行数据采集的小规模预演。试调查的目的在于评估调查工具、调查过程以及受访者选择程序。试调查采用小样本调查（常由少量访员开展），长时间以来，在调查研究中是一项标准实践。

历史地看，试调查会获得有关调查和调查问卷的两类信息。第一，“访员汇报 ”（interviewer debriefings），报告访员的看法，类似于访员的焦点小组讨论。在讨论中，访员报告在试调查中遇到的访题问题以及其他问题。通常，访员还会就如何让访问流畅以及

如何修订访题提供意见。第二，在试调查中，基于应答的定量信息。把试调查数据录入为数据表，调查设计者可以从总表中看出缺失值高的选项、不在选项范围内的值、与其他应答值不一致的应答。此外，对方差小的选项（大多数受访者的应答相同），则要么剔除，要么改写。

在试调查中录音，即对如何读出访题和应答访题作系统观察，也能对访题提供有用的信息（Oksenberg，Cannell，and Kalton，1991）。“行为编码 ”（behavior coding）就是对访员—受访者之间的互动进行系统分类和列举，用以刻画两者在问与答过程中的可观察行为。表8.1给出了问卷中一道访题的一次访问编码的示例。

表8.1　访员与受访者行为编码示例

编码分类	描　述
提问行为(单选)	1.访题读出准确。 2.访题读出有小变动。 3.访题读出且提示题意。
应答行为(可多选)	1.打断提问。 2.澄清访题。 3.给予充分应答。 4.给予合格且准确应答。 5.给予不充分应答。 6.回答“不知道”。 7.拒答。

运用如表8.1的编码对每一次访问进行编码。行为编码员考察的是，访员是否按照培训要求读出了访题，受访者有怎样的行为。经过

编码，每一次访问的问卷—应答行为就变成了数据集。接着，研究者就可以对每一道访题进行统计分析了，如：

1）按照要求读出访题的百分比。

2）受访者澄清访题某个维度的百分比。

3）受访者最初不作充分应答使得访员为获得可编码的应答而解释访题的百分比。

在问答过程中，有许多有意思的、可编码的维度。未来需要研究的一个重要领域是，识别在访问中可进行可靠编码且有助于改善数据质量的行为。

Oksenberg等论追问与行为编码

Oksenberg，Cannell和Kalton（1991）报道了一项用行为编码评估访题的研究。

研究设计：从既有的问卷中抽出一些访题，组成一份有60个选项的问卷，由6名电访访员进行测试。行为编码发现了访题的一些问题。之后，设计者进行修订，并再次进行了100个样本的测试。

研究发现：从下面的3道访题中，有了如下的行为编码：

（1）您去看医生或去医院的目的是什么？

（2）您最近一次看医生或去医院，自己花了多少钱？不包括保险已经出的或将要出的钱？如果您不知道确切数值，请您做最准确的估计。

（3）您上一次做一般体检是什么时间？

每道访题的问题率

访题	（1）	（2）	（3）
访员行为			
措辞微调	2	30	3
措辞大调	3	17	2
受访者行为			
打断	0	23	0
澄清	2	10	3
不充分应答	5	17	87
“不知道”	0	8	12

第（1）道访题，相对而言，没有什么问题。第（2）道访题在访员和受访者双方都产生了问题。第（3）道访题有歧义，“一般体检”题意不清。通过修改第（2）道访题，被打断、澄清、不充分应答的情况就减少了。

研究局限：没有说明如何解决问题。

研究意义：通过问答行为编码，可以探测访题的结构性问题。

在试调查中加入行为编码，只需要在受访者同意的条件下进行全程录音，并根据表8.1计算相关的百分比即可。相对于试调查，行为编码的价值在于对问卷的评价系统、客观、可重复。正如Fowler和Cannell（1996）报告的那样，当两位访员独立评估同一套访题时，两者用表8.1获得的行为比率的相关性达0.75～0.9，说明无论是哪位访员访问，出现问题比率的高低稳定一致。

8.6　随机或分组实验

有时候，设计者会采用实验来比较不同数据搜集方法、不同调查过程以及不同的问卷版本。这些实验，既可以是独立的研究，也可以作为试调查的一部分。如果在针对问卷、流程实验时采用随机样本，通常，实验也是这么做的，则被称为随机实验（randomized experiments）或折半实验（split-ballot experiments）。Tourangeau（2004）描述了这类设计中的一些议题，并引用了一些折半实验的例子。在我们的调查实例中，全国药物使用与健康调查（NSDUH）在运用折半实验上尤其活跃。这项调查已经进行过多项折半实验以检验不同数据搜集方法和访题措辞对药物滥用应答的影响（参见Tumer，Lessler，and Devore，1992；Tumer，Lessler，George，Hubbard，and Witt，1992；and，更多最近的例子，参见Lessler，Caspar，Penne，and Barker，2000）。相似的，当前人口调查（Current Population Survey，CPS）的问卷希望透彻了解失业问题，就拿旧版访题与新版访题进行了比较实验（Cohany，Polivka，

and Rothgeb，1994），并清晰地呈现了因访题变化所导致的月失业率变化（由CPS反映的）。

这类实验提供了不同措辞、访题顺序、数据搜集方式等方法特征影响应答的明确证据。不幸的是，尽管人们可以展现不同版本的工具或流程会获得不同的应答，折半实验并不能解决问题，不能说明哪个版本或流程能获得更好的应答。当用外部有效数据作为比照时，还会出现意外。如果有较强的理论依据，也可以说明哪个版本更好。譬如，Turner及其同事就认为自访问卷能改善药物滥用的应答，因为应答率上升了。一些早期的研究表明，受访者会低报药物滥用，报告率的提高就意味着应答的改善。Fowler（2004）描述了另一些折半实验，在这些实验中，尽管没有效度数据，似乎也说明有些版本的访题获得了更好的数据。

Fowler的例子，由于时间和成本的缘故，样本量较小（有的样本量小于100）。如果希望看到细微的差别，就需要较大的样本量。由于许多调查的预算并不宽裕，即使是小规模的折半实验，也会显得成本较高。正因为如此，在正式调查之前做折半实验，并不是常见的现象。尽管如此，实验方法还是提供了其他评估方法不曾有的、评估措辞变化对应答影响的途径。

8.7　运用提问标准

要识别的问题不同，运用的方法也不同。这一节，我们讨论在评估访题时，哪些方法能够满足评估访题的三条标准。

访题的“内容标准 ”（content standard）是指访题是否询问了要询问的事儿。故，要从两方面进行评估。第一，从研究者的角度，就要求访题必须搜集到研究目标所要求的信息。对此进行评估的唯一方法就是询问专家（研究者或其他专家），看访题提供了用于分析的信息没有。第二，受访者是否能如实提供信息。只有受访者在某种程度上提供准确信息，调查才能提供有用的信息。评估受访者在多大程度上能应答访题的基本方法就是焦点小组和认知访谈。通过焦点小组就能了解受访者知道什么。通过认知访谈，就能了解一组访题是否能顺利地获得应答，以及访谈是否提供了分析所需要的信息。

“认知标准 ”（cognitive standard），即受访者是否理解并应答访题，通常运用于认知测试，这也是认知访谈要做的事。此外，还有三种访题评估活动可以帮助识别访题的认知问题：

1）焦点小组可以识别对措辞理解的不一致性、概念歧义，以及受访者不能应答等问题。

2）专家评估可以识别有歧义的术语和概念、应答困难，通常会先于任何认知访谈。

3）试调查的行为编码可以识别不清晰的、应答有困难的访题。

“可用性 ”（usability）评估，是指调查工具是否可执行，也是试调查的基本目标。此外，在认知访谈之前，专家评估可以识别可能给访员和受访者带来困难的问题。在自访问卷中，可用性测试是最有价值的。在有控制的实验条件下或典型调查中，调查人员可以观察受访者的行为、对任务的理解以及完成任务的努力。借助于计算机辅

助（Couper，Hansen，and Sadowsky，1997；Hansen and Couper 2004；Tamai and Moore，2004），则可以记录键盘操作，观察对撤销键的使用，以及违规操作的比例。尽管实验不能捕捉实地调查的所有问题，却能够避免让实地调查变得更糟糕。

8.8 访题评估工具小结

这一章讨论的评估调查访题的所有技术各有长处，也有局限。下面，我们就此进行讨论。

1）专家对内容进行评估为数据的分析利用提供了数据使用者视角，却并没有告诉我们什么是最好的访题，即为了提供必要的信息而让受访者能够最准确地应答的访题。

2）问卷设计专家对问卷访题的系统评估至少是成本最小，也最容易操作的方法（Presser and Blair，1994；另参见文本框）。专家们或许在访题是否清晰、受访者应答是否有困难等问题上有分歧，那也只能用类似于认知测试的方法来让受访者实测才知道到底如何。也许我们可以把访题的问题清单列得更长，不过，专家评估的价值却不会因此增加多少。

3）焦点小组是一种有效的方法，6～10人就访题在实地调查中的问题进行讨论，可以获得主意和观念。不过，小组的形式并不利于了解某个受访者如何理解某道访题，以及如何应答。

4）认知测试是了解受访者理解访题以及如何应答的有用方法。不过，认知测试的人群规模很小，不一定就能够代表整个目标

总体。因此，对认知测试唯一的担心就是结果的代表性。我们不可能知道通过认知测试获得的问题或议题分布是不是可以代表目标总体。此外，实验室的有酬测试可以让受访者能够且愿意应答。最后，不同的认知测试者可能得到不同的结论，也许会盯着受访者提供证据（Beatty，2004）。另一个问题是，认知测试提供的访题问题并不系统，而只提供了测试数据（Conrad and Blair，1996）。

5）在实验室环境下的可用性测试的优缺点与认知测试类似。

6）试调查是在真实环境下测试工具和调查流程的最好方式。通过行为编码，研究者可以获得在真实环境下问答过程的系统信息。如果可以把结果数据化并向试调查访员进行简报，则更加有益。试调查的局限是，由于研究者试图去复制真实的调查程序，真正去追问访员和受访者如何理解并应答访题的机会不多。

一个核心的问题是，不同方法产生的结果在多大程度上提供了相似的信息。表8.2列出了不同研究中比较不同方法的结果。其中的第一项研究，Presser和Blair（1994）比较了试调查、专家评估、认知测试，以及试调查的行为编码（参见文本框 ）。他们发现，不同方法获得的结果有部分是相似的，专家评估和认知访谈倾向于发现更加综合的问题，试调查则更多地发现了选项的可用性问题。专家提供的问题最多，但不总是有用的问题。

表8.2　比较不同的访题评估方法

Presser and Blair, 1994

测试方法	标　准	结　论
1.传统试调查 2.行为编码 3.认知访谈 4.专家评估	• 发现问题的数量 • 发现问题的类型(即把问题分为4类) • 在测试中使用相同方法	1.传统试调查和行为编码发现最多的访员问题。 2.专家讨论和认知访谈发现最多分析性问题。 3.专家讨论和行为编码在测试者中都发现更多类型的问题。 4.行为编码最可靠,只是不能提供问题的原因,不能发现分析性问题,也不能区分受访者理解上有问题还是受访者在敷衍。 5.专家小组评估最有效率。 6.最常见的问题是理解受访者语义。

WiIlis, Schechter, and Whitaker, 1999

测试方法	标　准	结　论
1.认知访谈(访员在两个机构实施) 2.专家评估 3.行为编码	• 发现问题的数量 • 不同方法呈现问题的内外一致性(通过测量不同方法在时间选项分类问题上的相关性) • 发现问题的类型(基于5分法)	1.专家评估发现最多问题。 2.不同行为编码测试之间的相关性最高(0.79),接着是在两个机构的认知访谈(0.68)。 3.在两个机构实施的试调查大多为理解和沟通问题;在运用不同技术进行分类时,共识是需要做二次编码。

Rothgeb, WiIlis, and Forsyth, 2001

测试方法	标　准	结　论
三个研究机构测试了三种问卷,依据作者的分类标准对问题进行编码: 1.非正式专家评估 2.正式认知评估 3.认知访谈	• 发现问题的数量 • 基于对每项的综合评分(每个机构、每个方法都根据是否被标记为问题记0~9分)看不同方法的一致性	1.正式认知评估(QAS)发现问题的数量最多,给出的问题识别标记却很少。 2.非正式专家评估和认知访谈发现问题的数量相似,问题的类型却不相同。 3.在机构间发现的问题相似,用不同的方法发现的问题却不那么相似;机构间的综合评分有些一致(r 值为0.34~0.38)。 4.三种技术都能识别沟通和理解问题

续表

Forsyth, Rothgeb, and Willis, 2004 注：这是对 Rothgeb, 2001 的继续研究		
测试方法	标　准	结　论
1.非正式专家评估 2.正式认知评估（QAS） 3.认知访谈	• 在随机电话调查中采用随机实验方法，并控制问卷（2001年试调查的原始问卷）和实验问卷（依据试调查结果修订过的问卷）进行比较 • 根据行为编码数据和访员评估数据，把受访者和访员的问题区分为低、中、高	1.在试调查中被访员列为高的问题在实地执行中依然有许多问题（根据行为编码和访员评估）。 2.在试调查中被受访者列为高的问题在实地执行中依然有许多问题。 3.在试调查中需要回忆的以及敏感的访题，在实地执行中具有更高的无应答率。 4.在实验问卷中被修订过的访题并没有显著降低无应答，也没有减少在行为编码中发现的问题，却显著地减少了受访者的问题（根据访员评估）。不过，根据访员的评估，访员的访问问题似乎更多了。

DeMaio and Landreth, 2004		
测试方法	标　准	结　论
1.三种认知访谈方法（3个研究小组在3个机构用3种流程） 2.专家评估	• 识别问题的数量 • 识别问题的类型 • 识别问题的技术 • 不同方法/机构之间一致的频率	1.认知访谈的不同方法识别出了不同数量和类型的问题。 2.认知访谈小组比专家评估发现的问题少，尽管在3个机构中都识别出了“有缺陷”（至少有2名专家认为有问题）的访题。 3.认知访谈中发现的问题，在专家评估中也会被认为是问题。 4.不同小组运用了不同的追问方法。 5.对修改过的问卷再次进行认知访谈的结果表明，与修改前的问卷比较，只有一组发现的问题减少了。

Presser和Blair的结论经受住了时间段检验。譬如，最近Forsyth，Rothgeb和Willis（2004）的一项研究采用更加尖锐的视角看待不同方法对问题识别的重要性，发现（正如Presser和Blair做过的）了一些识别相同问题的方法，也发现了识别特定问题的方法。专家是最能发现问题的人，不过，其中的一些问题也许对数据质量没有什么影响。对数据质量的弱测量，是长期以来挑战访题设计和评估的

问题，也使得评估研究有结果却没有结论。我们不能确认发现的问题就真的会降低应答的效度。

毫无疑问的是，不同的技术之间互有补充。每一种技术都有长处和短处，提供的信息对应着不同的议题。如此，许多调查都要把试调查与不同的方法结合起来。到底使用什么技术，则要视调查的预算以及对访题的先期实验。采用新问卷且试调查预算很有限，就只能少采用焦点小组，进行一次专家评估、1～2轮认知访谈，以及小规模的试调查就可以了。专家评估和认知访谈的成本不高，两者也能揭示许多潜在问题（Presser and Blair，1994；Forsyth，Rothgeb，and Willis，2004）。焦点小组则可通过受访者的反馈来调整问卷使用的概念和术语。小规模试调查可以发现调查操作问题。如果使用已有的问卷且修改量较小，就可以不做焦点小组和试调查，认知访谈也可以专注于新访题。

另一个极端是，大规模的调查，如果不作大规模试调查是不可以实施的。如果不作充分的试调查，给正式调查带来的失败风险就会更大。例如，2000年的人口普查就进行了一系列实地实验，并把实验中积累的问题，放入1998年在三个区域（加州的萨克拉门托，威斯康星的梅诺米尼县，以及包括哥伦比亚、南加州在内的11个县）进行的预演调查，受访者数量达到几十万人。最近的一次NSDUH也对问卷进行了大规模的折半实验（参见Lessler，Caspar，Penne，and Barker，2000）。

不过，这些技术的一个主要问题是很少能说明识别出来的问题如何影响了数据质量。有一些证据表明识别出的问题对调查估计值有重要影响。譬如，Fowler（1992）的研究表明，如果受访者对访题关键词的理解不一致，就可能产生系统误差。Mangione，Fowler和

Louis（1992）及其同事的研究也表明。如果受访者感到访题很难，访员就不得不解释，以获得充分的应答。如此，访员就会影响应答，导致标准误的变动。无论如何，到目前为止所讨论的技术，加上折半实验可能出现的意外，都不能告诉我们某道访题可能出现什么类型以及多大程度的误差。此外，我们也缺乏（无论是独立的还是组合的）评估以寻求追问的最好技术，因此，需要不同的研究与分析。这就是下一节的内容。

8.9　在测量质量的概念与统计估计之间建立关联

非常不幸的是，在调查方法中用于表述测量质量的术语并不是标准化的。心理测量和抽样统计都有自己的术语。关注于个体应答者对访题应答的使用了“效度”和“信度”概念。关注于对个体应答进行统计归纳的使用了“偏差”和“方差”概念。

8.9.1　效度

“效度 ”（validity）在不同的学科有不同的用法，即使在调查研究中，不同的研究者也可能意指不同的事儿。“效度”，一般指对意图测量事物进行测量的精准程度。在不同的情境下，对这个定义的应用也有不同。不幸的是，这个定义并没有说明用于评估效度的方法。早期的含义（Lord and Novick，1968）基于测量过程的简单概念模型，即把测量看作接近概念真实含义的实现过程。也就是说，每次

调查测量理论上是可重复的，如此，给定受访者对给定访题的每一次应答就是接近概念真实的一次尝试。正如在第2.3节呈现过的，令

μ_i =第i 个应答者给出的真值。

Y_{it} =第i 个应答者在第t 次尝试中的应答。

ε_{it} =Y_i 在第t 次尝试中相应的应答与真值的离差。

则应答过程的概念模型如下。在调查中，μ 是第i 个应答者的尝试（被称之为第t 次尝试），而不是应答μ_i ，因此，应答者提供的应答实际为Y_{it} ，

$$Y_{it} = \mu_i + \varepsilon_{it}$$

在上面的等式中，由两个概念的相关性（应答者和尝试）来测量“效度”，即Y_i 与μ_i 的相关性。也就是说，一般而言，Y_i 越接近于μ_i ，则效度越高：

$$\text{效度}(Y) = \frac{\sum_{i,t}(Y_{it} - \bar{Y})(\mu_i - \bar{\mu})}{\sqrt{\sum_{i,t}(Y_{it} - \bar{Y})^2 \sum_i (\mu_i - \bar{\mu})^2}} = \text{相关}(Y_i, \mu_i)$$

故，效度由0.0到1.0之间的数值表示，数值越高，效度越高。后面我们会讨论在实践中效度估计的两种方式：运用外部数据和同一个调查的多个指标。

运用外部数据。来看之前我们讨论过的例子。

在过去的12个月中，自<u>（日期）</u>起，您找过医生或医生助理多少次聊您的健康状态？不包括您生病去医院的情况，此外的其他任何情况下您找过或聊过的，都算。

这是一道事实题。原则上，应答的质量取决于事实定义的准确性。在给定时间范围内，受访者可能有若干与题意相符的看医生行为，尽管在哪些算、哪些不算上可能有模糊之处（譬如受访者向医生电话咨询算不算），不过，也是可消除的模糊。就这道访题而言，如果受访者在过去的12个月中看过两次医生，则μ_i为2。如果Y_{it}也为2，那么这次尝试中，应答与真值之间就没有偏差。如果所有应答者反应的都差不多是真值，则效度为1.0。

在某些调查中，已有的记录或数据也可以用于评估调查应答的质量。假定第i个应答者的应答为μ_i，不管其真值是什么，都是可计算的。如果进一步假定调查是一次有代表性的尝试，则效度可计算如下：

$$\text{一次尝试的估计效度} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(\hat{\mu}_i - \bar{\hat{\mu}})^2}}$$

这里，

μ_i=记录中的变量值，有时候又称为研究的“金标准”。

$\bar{\mu}$=所有应答记录值的平均值。

即受访者应答值与记录值之间的相关性。如果相关性接近于1.0，则测量效度高。

多指标估计效度。第二类如态度，由于没有事实依据，因此也就无从确认受访者应答的准确性。

> 如果把国家的商业环境看作一个整体，您认为在未来的12个月，我们的投资环境将会是好还是坏，或其他？［消费者调查，SOC］

这道题希望测量经济预期，未来12个月的经济状况与应答的准确性没有直接关系。因此，判断受访者应答如何是没有事实标准的，即主观判断。大多数受访者对未来的经济状况并没有作过预判，而是在应答时通过参照其他情景给予应答。譬如，受访者也许通过失业率或股票市场趋势作出判断。因此，很难说依据什么进行的判断更加准确，而且，我们也没有针对任何依据的判断。在这种情况下，针对这道访题，对任何受访者而言，μ_i 是不清楚的，要测量效度就更加复杂了。

把调查应答与来自于其他的“金标准”进行比较是一种理想模式。不过，在调查研究中，找到准确的外部信息用于对应答进行评估并不常见。此外，当询问主观状态时，如知识、感受，或观点，几乎找不到独立于受访者应答之外的“金标准”。在缺乏外部标准的条件下，对应答效度的评估则有赖于以下三种分析之一：

1）把应答与其他在理论上具有效度的调查应答做相关分析。

2）如果应答用于测量潜在建构，则把理应具有差异的群组应答进行比较。

3）用不同的措辞或数据搜集方法，对可比较的样本应答进行比较。

第一种是最常见的评价效度的方法。譬如，如果希望测量某人的健康状态，与认为自己健康状态不怎么样的受访者比较，认为自己健康状态不错的受访者的感受也应该不错，在工作中较少缺勤，也能做更多的事情。对这类结果的分析评估，被称为*建构效度*（construct validity）（Cronbach and Meehl，1955）。如果研究者没有发现预期的关系，那么就有理由质疑对健康的测量。

再譬如，精神健康问卷（Mental Health Inventory Five Item Questionnaire，MHI-5）是广泛应用的用于测量当前心理状态的系列问卷之一（Stewart，Ware，Sherboume，and Wells，1992）：

这些访题询问的是，在过去4周时间里您的感受以及事情的进展。对每一道访题，选择一个与您的感受最接近的选项。在过去的4周里，您感受的程度如何？

（a）您觉得自己是一个幸福的人吗？

（b）您感到过无精打采和忧郁吗？

（c）您觉得自己是一个紧张的人吗？

（d）您感到过平淡和平静吗？

（e）您感到过如坠无底深渊且任何事儿都不能让您打起精神来？

应答分类包括：“始终”“大多数时间”“很多时间”“有时”“甚少时间”“从没”。这些访题询问的都是受访者的心理状态。对这6个应答类别，分别赋予不同的值（譬如1～6）且取正值，5道访题的总分就为5～30分。

把其他的指标当作标准，通过计算与其他测量的相关性，就可以看到MHI-5在多大程度上测量到了要测量的状态，即效度，有时也称之为“同时效度”（concurrent validity），因为两套工具在同时使用。Stewart，Ware，Sherboume和Wells（1992）发现，MHI-5与心理悲伤（distress）测量的相关系数为-0.94，与心理压抑（depression）测量的相关系数为-0.92，与焦虑（anxiety）测量的相关系数为-0.86，与积极影响（positive affect）的相关系数为+0.88，与感知认知功能（perceived cognitive functioning）测量的相关系数为+0.69，与从属感觉（feeling of belonging）测量的相关系数为+0.66。作者的结论是，这些相关系数与人们预期的方向与关联性一致，因此，MHI-5有很好的效度。

可以用复杂模型来评估效度，同时也可以通过对不同测量之间相关的模式与强度考察效度（参见Andrews，1984；Saris and Andrews，1991）。想想之前讨论的MHI-5，建构模型法刻画的是把5道访题的每一个应答作为反映不同效度水平 λ_a 相同的潜在建构 μ_i 。

$$Y_{ai}=\lambda_a\mu_i+\varepsilon_{ai}$$

注意，这个模型是基本误差模型$Y_i = \mu_i + \varepsilon_i$的一个变体。这里，等式刻画的不是一道访题的应答，而是许多访题的应答，每一道都有不同的α。对每道访题的应答是潜在建构μ_i的函数，并记为系数λ_α。

例如，表述MHI-5测量过程的简答模型可以是一幅路径图，如图8.1所示。最上面的圆圈代表潜在建构μ_i，从圆圈传出的箭头意味着作为λ_α函数的指标（$Y_{\alpha i}$）建构“因”的值。α等于1，2，3，4，或5，是箭头线上的数。每一个方框代表一道访题，共5道访题。

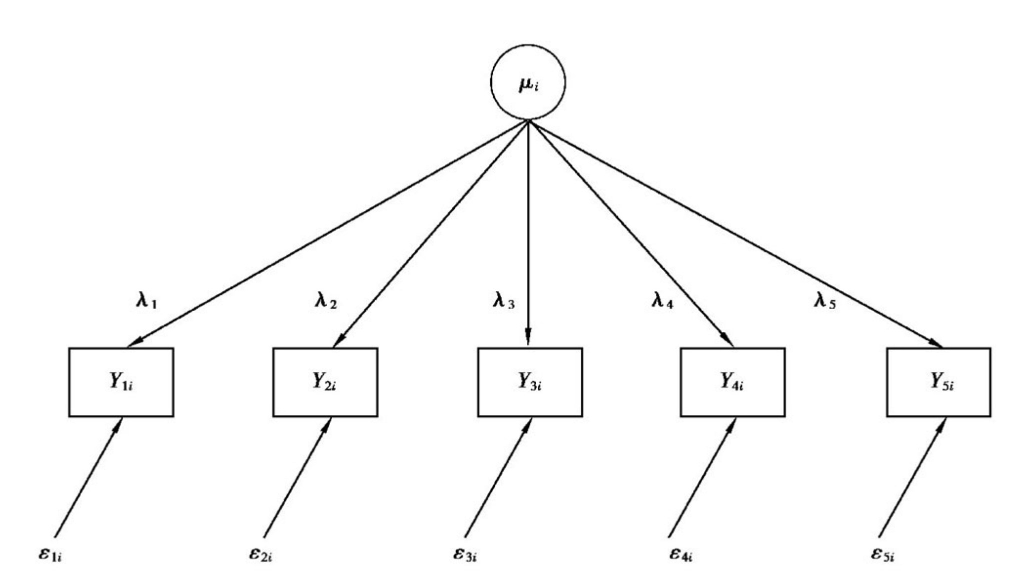


图8.1　表示$Y_{\alpha i} = \lambda_\alpha \mu_i + \varepsilon_{\alpha i}$的路径图，$\mu_i$的一个测量模型

简单地说，图示代表了5个等式。

$$Y_{1i} = \lambda_1 \mu_i + \varepsilon_{1i}$$

$$Y_{2i} = \lambda_2 \mu_i + \varepsilon_{2i}$$

$$Y_{3i} = \lambda_3 \mu_i + \varepsilon_{3i}$$

$$Y_{4i} = \lambda_4 \mu_i + \varepsilon_{4i}$$

$$Y_{5i} = \lambda_5 \mu_i + \varepsilon_{5i}$$

λ_a 是对每一道访题效度的测量，可以用多种方法进行估计（参见Andrews，1984；Saris and Andrews，1991；Saris and Gallhofer，2007）。

测量效度的另一种方法是比较不同组受访者的应答。理论上，不同组受访者的应答在潜在建构上应该不同。例如在美国，理论上，共和党人比民主党人更保守。因此，人们期待对测量保守性访题的应答，共和党人比民主党人应该得更高的分。这类评估严重依赖于不同变量之间关系的理论。在上面的例子中，如果共和党人和民主党人没有什么区别，那就意味着，要么访题对保守性的测量很差，要么共和党人和民主党人在保守性上没有那么大的差别。一般而言，在建构效度的评估中，很难区分到底是测量工具不好，还是理论有问题。

因此，对测量工具的评估只能在好理论的背景下进行。如果对误差来源摸不着头脑，那么当用两种方法搜集数据且得到了不同的结果时，就不能说哪个结果更准确或有效。

8.9.2　应答偏差

在涉及访题的误差中，最容易产生的混淆是“效度”如何与“偏差”有关。上面的讨论表明，效度是应答值与真值相关关系的函数。

因此，效度是某个受访者的应答属性。如果应答值与真值之间存在系统偏差，又会怎么样呢？例如，对第5.3.5节和第7.3.7节讨论的非社会期待行为就始终存在低报。在一些条件下，系统性低报可能不会降低应答值与真值的相关程度。例如，如果所有受访者都把自己的体重低报5磅，报告体重与实际体重的相关系数为1.0。然而，受访者的平均体重将比真实体重实实在在地少5磅。“应答偏差 ”（response bias）刻画的就是系统性的报告误差对如抽样均值统计量的影响。

折半设计可以用来检测偏差。Sudman和Bradburn（1982）提供了一个实例，就酒精饮料消费，比较了两道访题：

（a）在您饮酒的日子里，通常饮几次酒——一次、两次还是三次，或者更多次？

（b）在您饮酒的日子里，通常饮几次酒——一两次、三四次、五六次，还是七次抑或更多？

运用随机方法，指定受访者回答其中的一道访题。Sudman和Bradburn发现，那些应答访题（b）的受访者比应答访题（a）的受访者更可能报告饮酒次数不少于三次。由此，研究者认为，受访者倾向于低报饮酒量。在这种情景下，第二道访题的应答要比第一道更有效。值得注意的是，与前面讨论的效度测量相似，对偏差的测量也要求研究者对真值进行假设（如用系统记录与调查应答进行比较，就意味着假定系统记录没有误差）。

为了说明调查统计中的“偏差”，我们仍从相同的测量模型开始：

$$Y_{it} = \mu_i + \varepsilon_{it}$$

即，对访题应答值与真值之间有一个误差。如果误差项中存在系统性成分，那么期望值就不会为0。当调查观测到的期望值（$Y's$）与真值不同时，就意味着有“偏差”。即所有受访者的应答与真值之间存在系统离差：

$$偏差(Y_{it}) = \sum_t \left[\sum_i (Y_{it}) - \mu_i \right]$$

这一表达式与统计值（均值）相关，即所有应答者的平均值或期望值（表达式$\sum_i$部分）只是应答平均值。

也就是说，应答平均值是对真值平均值的有偏估计：

$$偏差(\overline{Y}) = \sum_i \left(\frac{\sum_i Y_{it}}{N} \right) - \frac{\sum_i \mu_i}{N}$$

上面表达式中的第一项是所有应答值的期望值。

尽管偏差的表达依靠μ_i值，效度的表达式则仅仅取决于μ_i与Y_i的相关性。简而言之，偏差概念因真值的存在而存在。至于知识、观点和感情等主观状况是否有真值，是有争论的。虽然心理计量学家曾经试图构建如态度等主观测量的真值，正如我们讨论过的，基于外部测量是不可能获得真值的。也就是说，对主观概念，只能比较两个或者更多受访者的报告值。因此，从技术上讲，偏差概念仅适用于在客观上可证实的事实或事件的测量。

实践中，有两种估计应答偏差的方法：运用调查之外的目标总体要素数据和运用不受调查误差测量影响的总体统计值。

运用目标总体要素数据 。 在不能确定存在真值的情况下，估计调查访题应答偏差的实际办法是与外部指标进行比较。例如，一些对调查偏差的研究把病历作为真值指标（Cannell，Marquis，and Laurent，1977；Edwards，Winn，and Collins，1996）。Cannell及其同事把受访者报告的某具体时间段内住院情况与医院记录进行比较。与之相似，Edwards和他的助手们也拿调查应答与病历进行比较来评估健康状况应答的质量（Edwards et al.，1994）。如果有医院记录可与调查应答进行比较，则研究者在评估应答质量就处在非常有利的位置。

让我们更细致地看看“记录核对 ”（record check）研究。在Cannell，Marquis和Laurent（1977）的研究中，入户调查的对象是在调查前一年内有人曾住院的家庭。研究者运用医院病历抽选要调查的家庭，观察病历记载的住院情况在健康调查中是否能得到如实反映。表8.3显示了按报告住院时间长短（病人住院的天数）及报告住院发生距调查时的周数的百分比。总体上说，受访者仅报告了住院情况的85%。然而，受访者的报告明显受住院时间长短和住院距调查时的时间远近的影响。距调查时越近的住院情况，受访者报告的越准确；对住院时间5天或以上情况的报告要好于短于5天的情况。

表8.3　没有报告已知住院的百分比（按住院时间长短和出院晚近划分）

出院时间	住院时间程度		
	1 天	2~4 天	5 天及以上
1~20 周	21%	5%	5%
21~40 周	27%	11%	7%
41~52 周	32%	34%	22%

数据来源：Cannell et al.，1977。

这个表是一个经典例证，说明随着时间的推移，哪些事件容易被忘记、哪些又还记得（参见图7.1 ）。表8.3意味着在统计估计中出现的偏差，一组受访者的住院平均数值。在调查前41～52周住院的平均数偏差比调查前1～20周住院的更大。

运用不受调查误差影响的总体统计值 。有时，尽管个体应答的准确性不明，还是可以用汇总数据来评估调查数据。例如，在1975年的一项赌博行为调查（Kallick-Kaufman，1979），有一道题询问受访者在合法情况下赌马时在每道下注的钱数。每道下注的总钱数都是公开数据。研究者虽然不能评估每位受访者应答的准确性，却可以对一年内调查下注钱数的估计值与公布值进行比较，以评估应答是否存在整体净偏差。研究发现，依据调查数据估计的结果与赛道公布的数据惊人地相似。由此，研究者认为受访者对赛道下注的应答既没有低报也没有高报（Kallick-Kaufman，1979）。

与之相似，选举日会公布投票结果，调查汇总的投票就可与公布的结果进行比较。即使没有个体应答数据，也可以依据公开的数据对汇总数据的偏差进行估计。

8.9.3　信度和简单应答方差

“信度 ”（reliability）是对重复测试中应答变异性的测量。信度强调的是受访者应答是否具有一致性与稳定性。因此，用方差进行表达，即所有受访者的应答变异性 ε_{it} 。可记为：

$$信度=\frac{\sum_{i}(\mu_i-\bar{\mu})^2}{\sum_{i}(\mu_i-\bar{\mu})^2+\sum_{i,t}(\varepsilon_{it}-\bar{\varepsilon})^2}=\frac{真值方差}{应答值方差}$$

如果应答离差的方差

$$\sum_{i,t}(\varepsilon_{it}-\varepsilon)^2$$

比较小，那么信度系数就接近1.0，对总体测量而言就意味着“高信度”。如果应答的变异性大（即应答离差的方差大），那么信度系数就接近于0.0。

在调查统计领域，习惯上不用“信度”术语，而是用“简单应答方差”（simple response variance）。我们可以把“简单应答方差”看作与“信度”相对的概念。如果对总体而言，某道访题的信度高时，其简单应答方差就小（在第9.3节中，我们会把“简单应答方差”与“相关应答方差”进行比较）。

O’Muircheartaigh论运用重访估计简单应答方差

O'Muircheartaigh（1991）运用当前人口调查（CPS）的重访数据对简单应答方差进行了估计。

研究设计：在CPS访谈后一周左右，不同的访员对1/18的样本对象进行了重访。符合要求的受访者在各种情况下都要报告数据。总差别率（GDR）仅是第一次和第二次报告之差。

研究发现：受访者自己的GDR比他人的大，如报告比他们年轻的人，或无户主家庭的报告情况。

研究局限：自报应答方差更大可能来自于未对受访者做随机化处理。报告人通常是经常在家的人，因此，也可能是失业者。没有测量应答偏差，也没有观察随时间变化的不稳定性。有时再访由更资深访员进行，且访谈方式不同于第一次访谈。

研究意义：研究确认了重访对应答稳定性的系统影响。显示了调查中与应答变化相关的再访数据。自报告的高不稳定性可能来自于就业状况的变化，也表明涉及快速变化属性的报告，且稳定性较低。

“信度”指的是无论在不同情形下还是同一测量的不同访题其测量的一致性。调查研究人员评估应答信度有两种主要方法：对同一受访者进行多次访谈和用不同指标测量同一个构建。

对同一受访者进行多次访谈 。重复访谈有时也称为“重访研究”（reinterview studies）。根据下面的假设，重访可以用来评估简单应答方差：

1）两次访谈没有看到构建的本质变化（即 μ_i 没有改变）。

2）测量的所有重要维度没有变化（即“关键调查方面”没有改变）。

3）第一次测量对第二次应答没有影响（即无记忆效应，第二次与第一次测量相互独立）。

尽管每一个假设都很复杂，在实践中，人们常用的方法是用同一道访题对同一受访者的两次访谈，并用所得应答的一致性来评估信度（例如，Forsmanand and Schreiner，1991；O'Muircheartaigh，1991）。运用重测，研究者从每一个受访者那里获得了两个应答值：Y_{i1} 和 Y_{i2} 。

运用重访研究，研究者可以计算不同的统计量以测量多次应答的一致性：

1）信度，如上定义。

2）不一致性指标（index of inconsistency），等于（1-可靠性）。

3）简单应答方差，即，

$$\frac{1}{2N}\sum_{i}(Y_{i1}-Y_{i2})^2$$

这里，Y_{i1} 和 Y_{i2} 分别指第一次访谈和再次访谈的应答值。

4）总差异率（gross difference rate），对一个二分变量来说，即是否是简单应答方差的2倍。

在这些测量中，并没有一个有突出的优势，因为它们都互为算术函数。不同的调查组织者在报告一致性测量时会依据自己的偏好而使用不同的指标。

表8.4给出了全国犯罪受害者调查（NCVS）报告不一致性的一些指标（参见Graham，1984），可以看作应答误差统计的实例。例如，报告锁或窗户受损的受访者比例的不一致性指标是0.146，对应的信度系数为（1-0.146）=0.854，即调查应答有高信度。值得注意的是，一般而言，报告可见证据或物品损失的不一致性较高（信度较低），而报告受到刑事犯罪袭击的不一致性却较低（不一致性指标=0.041，可靠性=0.959）。这似乎说明某类属性的访题更容易获得一致的应答。

表8.4　全国犯罪受害者调查（NCVS）中受害特征的不一致性指标

访题及其分类	不一致性指标	
	点估计	95%的置信区间
6c.有任何表明罪犯对建筑进行破坏的证据吗？（多选题）		
毁坏锁或窗子	0.146	0.094～0.228
破坏门或窗子	0.206	0.143～0.299
砸碎屏风或毁坏帐幕	0.274	0.164～0.457
其他	0.408	0.287～0.581
13f.拿走了什么？（多选题）		
仅拿了现金	0.276	0.194～0.392
女用小包	0.341	0.216～0.537
钱包	0.189	0.115～0.310
小汽车	0.200	0.127～0.315
小汽车的配件	0.145	0.110～0.191
其他	0.117	0.089～0.153
7d.那(伙)人撞您了吗？把您撞倒了吗？或者以什么方式侵害您了吗？（单选题）		
是	0.041	0.016～0.108
否	0.041	0.016～0.108

数据来源：Graham，1984，表59-60。

用不同指标测量同一构建 。评估信度另一个方法是在多道访题测量同一构建（如概念——译者注）。在主观状况测量中，这是常见的方法。其假设是：

1）所有访题都是同一构建的指标（即期望值相同）。

2）期望所有访题都有相同的应答离差（即简单应答方差或信度系数为常数）。

3）各项测量相互独立（即受访者对某道访题的应答不会影响其对其他访题的应答）。

科隆巴赫 α 系数 （Cronbach's alpha）是广泛用于多指标测量的信度测量（Cronbach，1991）。

第8.9.1节给出的五道访题MHI-5心理健康指数，就运用了MHI-5指标与其他心理健康指标的相关性比较来估计MHI-5的效度。科隆巴赫 α 系数测量的是综合多道访题测量的信度，且取决于访题数 k 及其相互相关的相关系数平均值 $\bar{\gamma}$ ：

$$\alpha = \frac{k\bar{r}}{1 + (k - 1)\bar{r}}$$

假定某道访题与其他访题的相关系数见表8.5。10个相关系数的平均值为0.539，则 α 值为0.85：

$$\alpha = \frac{k\bar{r}}{1 + (k - 1)\bar{r}} = \frac{5(0.539)}{1 + 4(0.539)} = 0.854$$

表8.5　MHI-5访题间相关系数示例

问　题	幸　福	情绪低落甚至沮丧	非常焦虑不安	平静甚至平和	沮丧到了极点
幸福	—				
情绪低落甚至沮丧	0.55	—			
非常焦虑不安	0.45	0.59	—		
平静甚至平和	0.62	0.51	0.54	—	
沮丧到了极点	0.49	0.63	0.56	0.45	—

科隆巴赫 α 系数越大，意味着信度越高或应答方差越低。令人遗憾的是，科隆巴赫 α 系数也表明，对某道访题的应答会影响其他的应答，进而产生高的正相关。如果 α 系数小，则意味着信度低以及访题没有真正测量相同的构建。

MHI-5的假设是：5道访题中的每一道都在测量同一构建的某个维度。如此，每道访题的应答既测量构建的共同部分，也测量构建的特定部分。譬如，幸福与平静都是好的情绪状态，却是不一样的情绪状态。而可以说明这5道题测量的是同一个构建的，则是每道题之间的高相关性。

并不是所有多访题测量的效度都是这样产生的。调查者可以自己定义某个复杂概念，选取项目的某个维度，且维度之间不一定以任何方式关联。这类测量的一个例子如“健康计划的消费者评估”（Consumer Assessment of Health Plans，CAHPS）调查，测量的是

病人在健康照料中的体验。在CAHPS中，运用了多组多访题测量，目的是归纳病人的体验，并把结果呈现给正在选择健康计划的人。其中的一组叫做“快速获得照料”，由4道访题组成，询问病人：（1）当他们呼叫医生办公室求助或求指导时得到回应的频率；（2）在日常照料中需要预约时得到响应的速度；（3）因病伤需要大夫的帮助得到响应的速度；（4）在预约时间的15分钟内看上病的频率。所有这些访题都与获得照料在概念上有关，不过，医生办公室对求助电话的应答好坏与病人候诊时间的长短之间并无关联。之所以把这些访题放在一起，是因为研究者将其放在了一起，而不是因为其测量的是同一个过程或现象。

即使访题之间是相关的，也不意味着其在某个特定层次也相关。上面例子的 α 系数也只有0.58。不过，复合测量作为一个整体，受访者对4道访题的应答对健康照料的感受也是非常好的预测指标（0.57），同时，对健康计划概念的测量也具有高信度（0.94）。因此，即使复合测量的访题并不总是测量同一个构建，也提供了有信度的测量，“获得快速照料”就是对受访者评价的重要预测工具。

简而言之，无论是用重复访谈还是用多访题来测量信度或简单应答方差，其背后的假设都是有争议的。然而，这两项技术在调查研究中经常使用，且实用性很强。

8.10 小结

评估调查访题有两方面内容：第一，确认是否问对了问题；第二，受访者是否按设想的去理解且没有意料之外的困难，以及在实践中是否容易实施。为此，我们给出解决问题的五种方法：专家评估、

焦点小组、认知测试、试调查和折半实验。几乎所有调查都在用其中的一种或多种方法开发问卷。不同的方法在某种程度上会获得不同的信息。一项调查到底要选择怎样的方法，则有赖于调查设计者关心的议题、调查预算以及访题的大部分或全部在之前是否使用过。可惜的是，这些技术是否真正来自于对应答影响最为严重部分，我们则知之甚少。

对调查访题进行统计评估测量的是应答效度或信度、应答方差和累加统计数据的偏差。评估效度有两种主要方法：把调查应答与外部资料（如各种记录）进行比较或判定调查测量是否与理论预期相符。

对信度和简单应答方差的评估，一般是对同一群受访者、用同样的访题进行两次访问，第一次在正式调查中进行，第二次在重访中进行。另一种方法是对同一个构建的多维度测量进行评估，检测不同维度访题应答的一致性。

把调查应答与外部数据进行比较，可以测量应答偏差。外部数据可以是目标总体的个体应答数据，也可以是累积统计数据。

把调查应答与精确的外部数据进行比较通常可以对调查误差提供有用的估计，只是，这类外部数据非常稀有。即使有，经常也是针对特定人群的，如特殊的HMO成员或特定医院的住院病人，这样的人群通常缺乏代表性。

越是使用间接方法评估效度，就越少能在测量层次提供误差的证据。尽管一次次地保证对访题的回答可以预测其他相关变量，但却不能提供测量误差水平的量化估计。不过，这些已经是我们能够评估调查访题回答效果的最好证据了。些术语有歧义，不同的人很容易有不

同的理解；有些访题则要求受访者回忆，会有困难。此外，您还怀疑不同的子群有不同的问题，譬如受教育程度较低的群体、少数族裔群体。在这种情况下，您会怎么做呢？

关键词

内容标准（content standards）

可用标准（usability standards）

焦点小组（focus group）

方案分析（protocol analysis）

访员汇报（interviewer debriefing）

随机实验（randomized experiment）

可用性（usability）

应答偏差（response bias）

信度（reliability）

重访研究（reinterview studies）

总差异率（gross difference rate）

认知标准（cognitive standards）

专家主题评估（subject matter experts）

认知访谈（cognitive interviewing）

试调查（pretest）

行为编码（behavior coding）

折半实验（split-ballot experiment）

效度（validity）

记录核对（record check）

简单应答方差（simple response variance）

不一致性指标（index of inconsistency）

科隆巴赫α系数（Cronbach's alpha）

进一步阅读资料

Alwin, D. (2007), *Margins of Error* , New York: Wiley.

Harkness, J., Vijver, F., and Mohler, P. (2002), *Cross-Cultural Survey Methods* , New York: Wiley.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E. (eds.) (2004), *Methods for*

Testing and Evaluating Survey Questionnaires, New York: Wiley.

Saris, W. and Gallhofer, I. (2007), *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, New York: Wiley.

Willis, G. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks, CA: Sage.

作业

---

1. 比较和对比问卷开发中三种方法的优势：认知测试、焦点小组以及在试调查中对访员和受访者行为进行编码。优势指的是，在正式调查之前，通过试调查获得问卷设计的缺陷并可以进行弥补。至少列出三种方法不同的三个优势。

2. 找一两个朋友对下列访题进行认知测试，想想在多大程度上符合认知标准。

（a）过去一年，您的收入是多少？

（b）您锻炼身体的频率—— 几乎每天，每周多次，每周一次，每月1次，或更少？

（c）您喜欢还是不喜欢美国的通用健康保险？

（d）在过去的一年，您用ATM机交易了多少次？

（e）想一想这个说法：“这几天我比平时更幸福。”您是强烈同意、同意、既不同意也不反对、反对，还是强烈反对？

3. 如果旨在测量客观事实，请列出两种方法，以检验访题获得的是有效度的数据。

4. 如果旨在测量主观状态如幸福感、焦虑感，请列出两种方法，以评估应答的效度。

5. 针对下面的每一道访题，设计两种用于认知访谈的追问方法，以发现访题任何可能的问题。

（a）在过去的四周，从［日期］到今天，您锻炼过吗？包括运动、体能训练、有氧活动，但不包括工作或家务中任何需要体能的活动？

（b）您每周摄取牛奶、黄油或其他奶制品多少次？

（c）在您现在的居住地，为满足基本消费，每月您和您的家庭收入最少需要多少钱（不包括任何抵扣）？

（d）有些人认为联邦政府应该降低通胀率，即使失业率会上升；另一些人则认为联邦政府应该降低失业率，即使通胀率会上升，您自己的看法呢？请在下面的7个刻度之中选一个适合您观点的点。

	1	2	3	4	5	6	7	
降低通胀率								降低失业率

（e）在过去的12个月，从［日期］算起，让您卧床至少半天的伤病有几天？包括要求您在医院留观的伤病。

6. 根据下面的方案，列出一种以发现问题为目的的试调查技术，并说明为什么采用这项技术。

（a）您正准备开发一个新问卷，为此，您需要了解目标总体对调查主题的所想所说，譬如他们的术语、界定等。此时，您会采用什么技术呢？

（b）您最关心的是在实地调查中访员与受访者之间的互动要尽可能地标准化。在非正式测试（譬如与合作者之间）中，您发现两者之间的互动有些尴尬，有时候在不得不打断应答以读完所有应答选项，有时候受访者又要您把访题重复一遍，还有时候，受访者要求您解释某个术语的意思（且问卷中没有标准定义）。如果希望在试调查中规范访员与受访者之间的互动，那么，最好要做什么？

（c）在问卷开发的最后阶段，您特别关注的是问卷涉及的理解和回忆问题。您希望做一次大规模的、历时几个月的实地演练。为了改善问卷，实演之前还必须做一些事情。譬如，有

7. 您用于做试调查的经费非常有限，且委托方认为他们准备的问卷已经可用。您看了一眼，发现了措辞、议题衔接等各种问题。此时，您只有做实地试调查的经费，而您有需要向委托方证明在试调查之前，问卷中有一些问题需要处理，且时间也非常有限。那么，在试调查之前，您会怎么做呢？

8. 您正在开发一个问卷，需要询问复杂的议题、长时段的回忆，且不同的子群（譬如受教育程度较低的群体、少数族裔群体）可能有不同的问题。委托方又催促着要尽快完成问卷开发，以进行长达

数月的实地演练，可问卷中的一些问题必须在实演开始之前处理好。您会怎么做呢？

9. 在您的机构中，有两位专家各自有自己擅长的主题。在问卷定稿阶段，只能有一位专家参与。两位专家都认为自己的版本是最好的。在正式调查之前，您安排了大规模实地试调查。在这种情况下，您将如何处理两位专家的要求？

10. 假设您在用一组多维度访题测量受访者过去一周的幸福感。这一组有4道访题，每一道都要求受访者用近义词描述自己，如开心、兴奋、乐观，4道题应答之间的相关系数为0.60。

    （a）这4道题的科隆巴赫系数是多少？

    （b）这个科隆巴赫系数又意味着什么？

11. 简述测量应答误差的两个方法，且说明各自的优势与劣势。

12. 您在做一项重访，一位新访员询问了从两周前刚刚结束的正式调查样本中抽取的受访者。有些访题应答的重复性很好，有些访题的应答与两周前获得的应答有明显不同。

    （a）描述两次应答之间吻合程度的概念是什么？

    （b）导致两次应答不同的可能的4个因数是什么？

    （c）如果应答非常不一致，用于测量应答效度的方法是什么？

    （d）如果应答高度一致，用于测量应答效度的方法又是什么？

9　调查访问

如果使用访员进行调查，则访员在调查过程中将扮演重要角色，对调查成本和数据质量也都有巨大的潜在影响。本章试图讨论访员会在哪些层面影响数据，以及研究者的设计选择如何减少或增加访员带来的调查数据误差。

9.1　访员的角色

访员在调查中扮演以下几种重要角色：

1）在使用区域概率抽样的住户调查中，运用列表名册编制抽样框。

2）在抽中的单位列出符合要求的受访者并根据规则选择受访者。

3）引导受访者配合访谈。

4）在调查互动中，帮助受访者扮演受访角色。

5）控制问答过程，提问访题，并在受访者回答不完全时进行追问。

6）记录受访者的应答，运用计算机录入数据。

7）校订应答记录的正确性，并将清理过的数据传送至调查组织中心。

其中，访员在抽样和获得受访者配合中的作用尤其重要。访员能否胜任这些职责，会对覆盖误差和无应答误差产生重要影响。在一般人群调查中，获得受访者合作是访员最困难的任务之一，因此也是访员培训和后续督导中非常重要的问题。这个问题，在前面章节中已经提及。本章将重点关注一旦受访者同意接受访问，访员如何影响数据的质量。

前面的章节区分了调查数据的两种误差：一种是与目标值的系统性离差或固定误差，即“偏差”（biases）；另一种是估计方差（variance of estimates），指概念化验证估计值的不稳定性。已有研究表明，在调查中访员会对这两种误差产生影响。

9.2　访员偏差

问卷应答的系统性访员效应来源，有以下三组发现：

1）与受访者自访问卷相比，访员在场会引起社会非期许行为的少报。

2）当访员的可观察特征与提问主题相关时，受访者会改变其应答。

3）改变回答与访员经验密切相关。

9.2.1 报告社会非期许的属性时，系统性访员效应

在第5章和第7章已经提到，许多研究比较了有访员调查和自访问卷的数据后发现，访员在场就会导致受访者应答偏差（例如，Turner，Forsyth，O'Rilly，Cooley，Smith，Rogers，and Miller，1998）。敏感行为（如违禁药物使用）访题似乎最容易受到访员在场的影响。对这种效应最好的理论解释是“社会在场”（social presence），访员的“社会在场”会促使受访者在应答时会考虑社会规范。迫于社会规范压力，受访者就会少报社会非期许性行为。这种效应是系统性的，且普遍存在，在误差结构中，被视为偏差。

9.2.2 涉及可观察的访员特征时，系统性访员效应

已有大量研究发现，在特定条件下，访员的可观察特征会对受访者行为产生影响。当访员的可观察特征被认为与提问主题相关时，这种影响经常发生。在这些情形中，访员的可观察特征可能影响受访者对问题的理解，受访者会依此判断何种答案更合适。

例如，Robinson和Rhode（1946）的一项研究表明，在反犹主义态度上，有常见犹太姓氏或显著犹太人特点的访员，会使受访者表达反犹主义态度或观念的比例降低。与此类似，在美国，涉及对待非裔美国人的态度，非洲裔访员会更少得到来自白人的敌意或畏惧应答（Schuman and Hatchet，1976）。Schuman和Converse（1971，参见

文本框 ）指出，许多涉及种族关系的访问其实并不受访员种族的影响，但与对黑人态度密切相关的访问，受访者的应答则常常因访员的肤色而不同。一项常被提及的研究表明，青少年会根据访员的年龄而给出不同的应答（Ehrlich & Riesman，1961）。接受救济的母亲面对远道而来、陌生的中产阶级访员报告的收入水平，通常会比她们在与自己的人口学特征相近的访员面前的报告更准确（Weiss，1968）。男性和女性面对不同性别的访员也会报告不同的性别态度（Kane and Macaulay，1993）。

Schuman和Converse论美国访员的种族效应

1968年，美国种族关系高度紧张的时期，Schuman和Converse（1971）在底特律和密歇根做了一项访员种族效应的研究。

研究设计：25名专业黑人访员与17名白人大学生访员分别进行了330次和165访谈，调查对象是黑人户主和他们的配偶。对不同种族的访员随机分配受访样本，每名访员负责5户。130道访题涉及种族态度、工作和生活经历、背景等变量。

研究发现：大部分（74%）访题的答案并没有因访员种族而表现出差异。然而，对种族态度类访题的应答显示，黑人访员比白人访员让更多受访者给出诸如敌视白人的应答。社会经济地位越低的黑人，效应越明显。

给定访员种族类别时某一特定答案的百分比

问　题	类　别	类别百分比	
		白人访员	黑人访员
您认为可以信任白人吗？	信任大多数白人	35%	7%
黑人父母最适合与黑人教师相处吗？	是	14%	29%
您最喜爱的娱乐艺人？	只说出黑人明星的名字	16%	43%

研究局限：访员种族效应可能受访员年龄和访谈经验的干扰。本研究并没有检验访员族群内部的差异性。

研究意义：当调查涉及种族内容时，研究者要考虑访员种族的因素，有可能会影响受访者感知到的问题意图。

每个社会的访员都存在一些可观察的外表特征（如声音、外表，行为倾向），这些特征都有其社会意义。在美国，这种可观察的特征表现为种族特征；在魁北克，则表现为访员的方言（Fellegi，1964）。当被问及与访员外表特征有关的访题时，受访者倾向于借助这些外表特征去选择应答。为了照顾访员特征而做应答往往是有偏差的。如果访题与访员的外表特征无关，则不存在这种影响。

以上的研究都涉及观点和态度。如前所述，在涉及主观现象时，偏差概念非常复杂，甚至可能意义不大。这里我们想指出的是，在某

些情况下，受访者的应答会由于访员特征而存在系统性差异。

另一方面，这些研究最值得注意的是，许多访题并不会受到访员可观察特征的影响。似乎在大多数情况下，访员都成功构造了一个与其可观察特征无关的访谈语境。只有在访题直接且特别指向访员明显的可观察特征时，这种语境才会有系统性影响，并对数据产生作用。因此，如果研究者想要测量访员的种族、性别、年龄之类的特征对态度或行为应答的影响，就要考虑如何派遣访员。在许多情形中，我们并不能确定与受访者相似的访员是否一定会造成数据偏差。即使进行的是主观测量，数据质量的构成也没有规定标准。面对这种模糊性，最好的应对方法是随机分派受访者和访员，即使不能减少与访员有关的误差，却可让测量这类误差成为可能。

9.2.3　与访员经验有关的系统性访员效应

有些混合证据表明，访员经验对调查过程会产生影响。虽然有经验的访员常比没经验的访员有更高的应答率，但却很难知道这种高应答率究竟是访员经验的影响，还是仅仅反映了一种事实，即难以得到受访者配合的访员没有再坚持下去。同时，已有证据表明，有经验的访员在准确读出访题并遵循访员规则上，要比新手访员表现得更不认真（Bradburn，Sudman，and Associates，1979；Gfroerer，Eyerman，and Chromy，2002）。

例如，在全国药物使用和健康调查（NSDUH）中，在是否使用过违禁药物询问中，有经验的访员得到的肯定应答要比没有经验的访员

少。表9.1显示了早期NSDUH使用自访问卷报告的药物使用情况。总地来说，较之于在1998年调查之前有过一次及以上NSDUH调查经验的访员，面对无经验的访员，报告使用过非法药物的受访者人数普遍高出21%。在表中，可以横向比较1998年两组访员在访谈前期（从第1次到第19次访问）和后期（第100次访问及以后）之间的不同。每组访员都按照访谈顺序分配一定工作量，有经验的访员在违禁药物使用上得到的肯定应答反而更少。在给定的工作量中，随着有经验的访员逐渐完成其调查任务，得到的报告数量也存在下降的趋势。

表9.1　分访员经验的调查顺序与受访者报告使用非法药物情况的百分比（1998年NSDUH）

	生命历程使用过非法药物的回答占比		
依照访问次数的访员分组	(1) 没有 NSDUH 经验的访员	(2) 有 NSDUH 经验的访员	比率(1)/(2)
1~19	40.9%	35.5%	1.15
20~39	38.7%	32.6%	1.19
20~59	38.2%	33.7%	1.13
60~99	39.0%	32.1%	1.21
100+	43.2%	31.7%	1.36
总计	40.1%	33.1%	1.21
比率 (1-19)(100+)	1.02	1.07	

数据来源：Hughes，Chromy，Giacoletti，and Odom（2002，表8.1）

为什么有经验的访员反而得到这种结果呢？请注意，NSDHU是自访问卷，访员不需要提问！进一步的多变量分析表明，有无经验的访员在应答结果上呈现出的不同，并非由于他们的样本人群特征不同。一种可能是，有经验的访员行为向受访者传递了一种信息，默认和容许他们可以仓促潦草地完成问卷。如果他们勾选了使用过非法药物，就不得不继续回答后面的追问访题。不勾选则可以减少完成问卷测量的

时间（补充培训和计算机辅助自访技术可能会抑制这类访员经验的影响）。

一项较早的研究曾使用医院记录来测量调查应答中的偏差（假定记录中没有错误）。在调查中，要求每个访员完成的访谈数量不同。将每个访员完成的工作量列表之后，对比受访者住院记录，发现二者之间存在很强的相关性（0.72）。完成访谈量最多的访员得到的受访者报告，比只有少量调查任务的访员更糟糕（Cannell，Marquis，and Laurent，1977）。最近的另一项研究也表明，访员执行访谈调查的次数越多，得到心理健康病症的报告就越少（Matshinger，Bernert，and Angermeyer，2005）。

这些系统性的访员经验效应，其机制是什么，还需要进一步的研究去发现。但最可能的情形是，访员经验效应是访员报酬体系的意外后果。对面访访员的持续反馈和评估，通常包含应答率和效率两项指标，缺乏对访员调查质量的反馈和调整，而调查质量会影响测量误差。有鉴于此，可以假设有经验的访员重视受访者的配合与效率，却忽视了激励受访者给出优质的回答。如果这一假设成立，可以预计在集中电话调查数据中会减少这种效应；并且在未来的研究中，要注重通过模式比较去挖掘其内在机制。

9.3　访员方差

在许多研究中，尽管上述提到的系统访员效应是可预测且反复出现的，在调查中依然是不同的访员对应答有不同的影响。在有访员参与的调查统计中，我们用“*访员方差* ”（interviewer variance）来描述数据整体的变异。与抽样方差（sampling variance）类似，访员

方差是一种概念模型，即调查只是对众多可能情况的一种认识。在不同的调查中会用到不同的访员。如果不同的访员会导致不同的应答，调查估计的值也会因此而不同。

至于访员如何增大调查估计量的变异性，这其中的多数逻辑类似于我们在4.4节讨论的整群抽样中的分群（clustering）对抽样方差的影响。在简单随机抽样中对标准误的计算，实际假设了每次观察、访问或回访看到的都是一个独立整体。在整群情形下，如果一群人非常相似，其相似度大于总体平均水平，那么群内的额外观察就不能得到与群外观察一致的信息，因为这不是对总体的重新独立观察。同样，如果访员影响应答，那么同一位访员进行多次调查就不能得到本应有的独立的新信息，而由另一位访员进行调查或不影响应答的访员多次调查却能够得到独立数据。

9.3.1　估计访员方差的随机性要求

测量访员方差的一个实际问题是对不同访员在受访者群体中得到不同应答的平均状况有两种可能的解释：

1）访员影响了受访者应答。

2）受访者为不同的访员准备了不同的应答。

为了估计纯粹由访员对受访者造成的影响，需要首先排除受访者之间真实差异的影响。“交叉样本分配 ”（interpenetrated sample assignments）是指分配给每个访员的样本是全部样本的概率子样本。交叉的工作量可以在排除受访者真实特征差异干扰的情况下测量访员

方差。这种设计给每位访员都分配一个随机子样本任务量。因此，如果不出意外，每个工作量应该得到相同的统计量估计值。显然，由于每个工作量是相对小的样本，由此计算的统计量会有较大的抽样变异性。然而，如果由访员工作量带来的统计量差异大于由抽样带来的统计量差异，就可以证明样本交叉分配中存在访员方差。访员对估计量方差影响的计算方法与整群样本影响的计算方法相似。

交叉的特征必须与要估计的访员差异性来源相对应。例如，在小规模城市社区面访问卷调查中，分配给访员的工作量通常需要他们跑遍城市的每个角落。为了研究这种设计下的访员方差，就可能为访员分配整个样本的简单随机子样本。这种设计要测量的是所有访员的访员方差。在全国性入户面访问卷调查中，一般会从初级抽样区域的居民中聘用访员。因此，每位访员通常只被分配到一个初级抽样单位（PSU）。在这样的调查中，对同一PSU的访员来讲，交叉设计就是将样本（住房单位群）进行随机分割然后分配给不同的访员，通常会将一个PSU随机分解为两部分，每位访员随机分得一部分。这种设计可以测量同一初级抽样单位的访员方差。在集中的电访中心，访员通常轮班工作（例如，星期一，星期四，星期六，下午3:00—9:00），可以将一个时点的活跃样本分配到轮班工作的所有访员。交叉分配则可以为同一班次的访员安排这段时间内活跃的一组随机样本。这种设计可以测量同一班次的访员方差。因此，交叉设计用来（即用调查应答）测量同一组样本的访员方差。例如，在全国性面访问卷调查中，绝不会给居住在纽约的访员分配洛杉矶的样本，交叉设计不允许这种样本分配方式。

如果样本分配不是交叉的，访员方差的估计就可能有干扰，即不同访员的差异会受到样本差异的干扰。例如，有的访员可能分到了一

些不同程度不愿接受调查的样本，有的访员可能仅分到了特定语言或种族的人群。这些偏差是否会影响访员方差，现在还不知道。只是，如果是给不同的访员同样的受访者，才可以对访员方差进行估计。

9.3.2 估计访员方差

对访员方差的估计需要用到测量模型。有两种基本测量模型：一种由Hansen，Hurwitz和Bershad（1961）提出，另一种由Kish（1962）提出。Kish模型更简单常用，这里介绍Kish模型。从前面讨论过的测量模型开始，$y_i=\mu_i+\varepsilon_i$，其中，μ 是真值，ε 是答案对真值的偏离，内生于应答y。当访员就某一特征 μ 进行提问时，受访者给出的应答常常偏离真值，这种偏差包括当第j 个访员提问时所有受访者普遍存在的偏差和第i 个受访者的特殊偏差。对于第i 个受访者，被第j 个访员提问访题 μ_i，答案是

$$y_{ij}=\mu_i+b_j+\varepsilon_{ij}$$

或

报告值=真值+访员偏差+受访者偏差

式中，b_j 是由访员产生的系统误差，是第i 个受访者对真值u_i 的偏离。此外，模型中有一个随机误差项，独立于访员效应，记为 ε_{ij}，是与接受第j 个访员提问的受访者的额外偏差。此模型也可用于估计访员偏差，在这种情况下b_j 的期望值不能为0（也就是偏差值不能

为0）。在研究访员方差的文献中，最常用的假设是b_j的期望值为0。

当b_j在不同访员间表现出差异性时，就会产生访员方差。一种测量方法是估计同一访员访问的所有受访者的总偏差。为此，可以使用组内相关系数（用于测量同一访员访问的所有受访者应答偏差的相关性）。

kish（1962）、Biemer和Stokes（1991）、Groves（1989）对估计组内相关系数ρ_{int}的方法及其复杂性做了大量研究。Kish（1962）方法的优点在于使用了简单方差分析的估计量。估计ρ_{int}的基本方法是：

$$\rho_{\text{int}} = \frac{\left(\dfrac{V_a - V_b}{m}\right)}{\left(\dfrac{V_a - V_b}{m}\right) + V_b}$$

式中，V_a指访员作为影响因素的单因素方差分析中的组间均方和；V_b指方差分析中的组内均方和；m指由一个访员完成的访谈数量。

Kish论访员方差

Kish（1962）通过对一项面访调查的研究，创建了一个简单的访员方差估计量。

研究设计：20个男性访员在接受一周培训后完成一项关于汽车制造工人的调查。问卷涉及工人对工长、督工、工会、经理和其他工作问题的态度。通过分层随机抽样将426名受访工人分配给访员，随后访员入户调查。用单因素方差分析（访员作为影响因素）的方法计算ρ_{int}值。

研究发现：46个变量的ρ_{int}值分布在0.00～0.09。意味着每个访员在平均工作量为23的情况下，访员方差的设计效应位于1.0～3.0。下表显示，有些变量的访员变异性较大。

运用模型$1+\rho_{int}(m-1)$得到的设计效应，总样本访员方差要高于子样本（譬如工厂的新工人）。这说明，每一个访员在子样本中的访谈数量要小于在总样本中的。

ρ_{int}值较高的统计量

问题/统计量	ρ_{int}	$deff_{int}$
你认为你应该挣多少?	0.092	3.02
1个以上不喜欢公司的理由	0.081	2.78
2个以下不喜欢目前工作的理由	0.068	2.50
对工作的批评意见的总数	0.063	2.36

研究局限：研究中只有20个访员，由此计算的ρ_{int}值极不稳定（在95%的置信区间内ρ_{int}值全部大于或等于0.03，排除了ρ

$_{int}$ 值为0.00的情况）。此调查也不是典型的住户调查。

研究意义：这项研究表明，调查估计不稳定性的大幅增加可能是因为访员效应，也有许多变量对访员的影响并不敏感。

ρ_{int} 的期望值位于-1/（m -1）～1.0，在访员和样本较少的情况下，有时会出现估计值略小于0的情况。接近0的估计值意味着受访者对特定访题的应答不受访员影响；据此认为，不同访员得到了相同的结果（在抽样波动的情况下）。ρ_{int} 值较大，说明访员对应答的影响较大，与访员有关的误差统计值也相应较高。虽然 ρ_{int} 值与分配给不同访员的受访者数量无关，但 ρ_{int} 值的意义在于：与访员有关的统计量的总方差与每个访员完成访问的平均数成比例。求解的等式是

$$访员设计效应 = deff_{int} = 1 + \rho_{int}(m-1)$$

式中，$deff_{int}$ 是访员差异性导致方差增加的程度；m 是每个访员完成访谈的平均数。

这种计算方式只适用于运用随机方法给访员分配受访者，以确保访员影响与不同访员分到的样本特征无关。

给定一个简单随机样本，$deff_{int}$ 是来自于访员差异性导致的简单随机样本均值y 的抽样方差的变动。例如，ρ_{int} 是0.02，每个访

员完成的访问量是101，那么$deff_{int}$ =1+0.02（101-1）=3.00。意味着样本均值方差增加了300%，或者说样本均值的标准误增加了73%。也意味着增加的置信区间为73%($\sqrt{3}=1.73$) 。

Groves（1989）对许多研究中ρ_{int}值的可计算性进行了鉴别，利用这些研究中的全部访题，计算出ρ_{int}值约为0.01。如果每个访员完成41次访谈，把ρ_{int}值转化为$deff_{int}$，其值为1.4，说明样本均值的方差增加了40%($\sqrt{1.4}=1.18$，或者说样本均值的标准误增加了18%）。在工作量很大的调查中，比如说每个访员完成100次访谈，$deff_{int}$值就是2.0，导致标准误增加41%($\sqrt{2}=1.41$) 。因此，即便ρ_{int}值较小，在访员工作量很大时，也会增加样本统计量的方差。

在任一调查中，访员对应答的影响也因访题而异。要求访员脱离脚本提问的访题似乎更容易受到访员的影响（如追问率，开放性应答的字数）。此外，有些受访者类型似乎对访员效应更敏感（譬如老年人）。如上所述，在模型中任意给定ρ_{int}值对调查估计的准确度严重受限于每个访员完成访谈的平均数量。

9.4　减少访员偏差的策略

减少访员偏差的实用工具，如果适用于所有访员，也一定能够减少访员方差。因此，简单区分某个工具适用于访员偏差还是方差，实际上是一种武断的做法。这一部分，我们将介绍为减少调查估计偏差而采取的干预访员行为的手段。

9.4.1　访员对激励受访者的作用

一项调查开始之初，大多数受访者其实并不清楚该如何表现。受访者该如何表现，其实是由访员界定的。第9.2.3节讨论了针对违禁药物使用、已知医院记录以及心理健康病症的访题，随着访员经验和访问规模的增加，受访者的报告数会下降。对此，一项较强的假设是，有经验的和工作量大的访员会向受访者交代更少的行为要求。

另一项研究的解释更加明确。这项研究搜集了调查对象一年前详细的住院治疗记录。数据质量就是受访者报告的住院情况占已有记录的百分比。访员用两种方式搜集受访者住院治疗的数据。对一半受访者，访员进行完整的健康调查访谈，包括询问住院治疗的情况。对另一半受访者，由同一批访员完成大部分提问，并在离开时留给受访者一份住院治疗情况的自访表格，要求他们填完后寄回。研究者在计算受访者报告的住院情况百分比时发现，无论是通过访员面访还是受访者自访得到的数据，受访者的回答质量与谁执行此次调查高度相关。如果一位访员的面访对象向他报告了较高的住院百分比，则受访者自访报告的百分比也会比较高（Cannell and Fowler，1964）。

这些研究都有力地说明，访员在一定程度上起着重要的激励作用，有效激励受访者行为的访员往往会得到更高质量的答案。Fowler and Mangione（1990）的研究揭示了产生这种结果的机制。在受访者接受健康调查的第二天，再次对他们进行访谈，了解其参与调查的经历。其中一个问题是，他们认为访员在调查时想要“准确答案”还是“仅仅是大致想法”。同时，让访员填写一份工作任务优先级问卷，询问访员“对准确性重视”还是“对效率重视”。数据分析显示，认为访员更想要“准确答案”而非“大致想法”的受访者，其应答质量

指标更可能得高分。更重要的是，受访者的认知与访员报告的优先级显著相关。对于将答案准确性作为最高优先级的访员，受访者更可能认为其在访谈中想要的是“准确答案”。相反，如果访员将效率置于最高的优先级，则受访者更可能认为访员只是想要“大致想法”。

这些研究表明，访员对受访者的应答起着关键作用。反过来，受访者对访员预期的认知则会对数据质量产生显著影响。可见，访员差异性对数据偏差产生影响的一个重要来源，就是访员能否与受访者进行高标准的沟通。

9.4.2 改变访员的行为

在这些工作基础上，Charles Cannell进行了一系列研究，试图通过统一对受访者的要求来减少访员之间的差异。在研究中，Cannell尝试了5种可能的方法来标准化访员传达给受访者的信息（Cannell，Marquis，and Laurent，1977）。

第一项实验是放慢访员讲话的速度。初步研究表明，访员的讲话速度可能与受访者认为应该如何应对访谈相关。如果访员的讲话速度非常快，受访者就会推测，访员优先考虑的是赶紧走完访谈流程。Cannell强调了访员放慢讲话速度的重要性，并且对放慢访员阅读访题的速度进行了试验。只是，目前并没有明确证实这种干预对数据质量一定会有积极的影响。

然而，模仿良好行为的实验似乎更有效。Henson让访员在访谈开始时播放一段录音，告诉访员在提问时要放慢讲话速度，提醒受访者在应答时要特别小心（Henson，Roth，and Cannell，1977）。听过录

音的受访者给出的应答，似乎在某些方面要比没听过录音的应答质量更高。

“系统强化”显示了更有效的结果。在这些实验中，Cannell要求访员，如果受访者的表现与准确应答的行为要求一致时，如要求访员澄清访题、耐心应答、充分且完整应答，则给予受访者积极肯定的反馈。相反，当受访者不假思索地给出应答时，访员就给予受访者负面反馈，鼓励他们更认真地思考访题。在一些实验中，强化良好的行为似乎可以提高数据的质量。

“程序化说明”为标准化访员向受访者传递行为期待信息提供了更简便、直接的方法。在这类实验中，访员要读一段标准化的说明语，向受访者强调耐心应答、思考应答和提供准确数据的重要性。在访谈过程中，定时向受访者传达这类信息。有证据表明，当访员阅读这些说明时，数据质量会显著提高。

“要求受访者做出承诺”可能是Cannell实验中最具创新特色的方法。在这项研究中，受访者同意接受访谈并应答少量访题之后，访员中断访谈，要求受访者签署一份承诺书。承诺书要求受访者尽其所能，提供准确的信息。受访者会被告知，如果不在承诺书上签字，访谈就无法完成。

最初尝试这个方法时，有人担心受访者会拒签，应答率也会迅速降低。然而，事实并非如此。几乎所有受访者都同意签署承诺书，且有证据表明，有承诺的受访者的数据质量远远好于没有被要求作出承诺的受访者。

这些干预技术从两个层面对受访者产生影响。第一，受访者改变对访员行为期望的理解。第二，产生了通过准确应答而表现更好的意愿。对受访者行为设立高标准的访员会得到高质量的数据，也就是说，与不交代行为标准的比较，交代了行为标准的访问获得的低报更少，系统误差更小。这些努力试图标准化访员行为，以使访员始终向受访者传达高标准的行为预期，也是减少访员偏差的有效策略。Cannell，Groves，Magilavy，Mathiowetz和Miller（1987）还把这些技术整合到一项电话调查中，将结果与全国健康状况调查（National Health Interview Survey）面访数据进行比较，发现当使用干预技术时，会得到受访者更多的应答。Miller和Cannell（1977）指出，这些技术的使用也可能导致某些事项或行为的少报，不过这样的少报反而是准确的。例如，在运用干预技术的实验组中，受访者相比控制组更可能回答他们会多看电视、少读书。

虽然有大量证据表明，Cannell发展的一系列干预技术有建设性的意义，但在常规数据搜集中却很少使用这些技术。一方面，实施这些技术要额外花费时间，虽然并不是太多。另一方面，在一般的调查中，应答误差也不容易发现。因此，对大多数调查者和数据使用者而言，这些为改善数据质量而进行的努力并没有非常明显的价值。尽管如此，访员引导受访者完成调查任务的方式还是会影响到数据的质量。当我们研究访谈调查并试图最小化与访员有关的误差时，这一核心观念理应得到更多关注。

9.5 降低访员方差的策略

为了使访员的访问尽可能一致，减少访员个体对访问结果的影响，研究者一般可以从以下三个方面进行控制：让访员提问的访题，让访员使用的程序和管理访员的方式。

减少访员方差最有效的方法之一是，制作的访题不要求访员在面对受访者时改变自己的行为，如澄清问题和追问不充分的应答。

此外，在数据采集的规范过程中，有五个方面的访员行为十分重要：

1）与受访者的互动方式是专业的、任务导向的，并最大限度地减少受访者附和或推测访题偏好的可能。

2）根据问卷逐字逐句阅读访题。

3）向受访者解释调查程序和问答过程。

4）无指向性地追问；也就是说，与其他应答相比，追问不会增加某一应答被选择的可能性。

5）记录受访者给出的应答，对受访者未说出的内容不加以解释和推测。

9.5.1　最大限度地减少要求访员非标准化行为的访题

访题对访员与受访者之间的互动有着可预测的、可靠的影响。最有力的证据是马萨诸塞州立大学、波士顿大学和密歇根大学进行的一项实验，即使用同一工具并行管理访谈参与者（Fowler and Cannell，1996），在受访者同意的前提下对访谈过程录音，并根据访题对访员和受访者行为逐一进行编码。结果表明，访员逐字阅读访题的比例与访题高度相关，而且受访者打断提问的比例、回答不充分的比例和需要追问的比例也与访题高度相关。这些结果意味着少数特殊访题对访员阅读和受访者应答有高度可预测的影响。有些访题很可能被误读，有些则不会；有些访题受访者会立即应答，不需要更多追问，有些访题则需要访员通过澄清和追问来获得充分应答。

Mangione，Fowler和Louis（1992）在一篇论文中阐述了这一发现的重要性。他们研究了不同访题的访员行为与访员相关误差之间的关系。最重要的发现是，越是需要访员通过追问来获得充分应答，访员对调查结果的影响就越大；第二个发现是，需要通过叙述形式应答的访题，访员追问的可能性更大，同时也对访员记录应答构成挑战，因而这些访题可能会比一般访题更容易受访员的影响。总体来说，这些研究都表明，设计一个好访题是减少访员相关误差的重要方式之一。在这里，“好访题”的标准是能最大限度地减少访员通过追问来获得充分应答。因此，其特征应包括：

1）访题措辞明确，让受访者在第一次阅读时即可理解。

2）应答方式明确，许多访题需要追问的一个重要原因是没有告知受访者该如何应答。

例如：

访题：“就个人健康问题您最近一次去看医生是什么时候？”

评论：这道访题未明确指出受访者该应答什么。

可能的应答：

“去年6月。”

“大约4个月前。”

“我怀孕之后。”

对上面的访题，这些都是合理的应答，在技术上都满足访题规定的要求。然而，为了方便分析，受访者应以相同的形式应答。这样，访员就不得不进行干预，向受访者解释在这三种可能的应答中，只有一种是调查所需要的。

下面这道访题也没有明确指出应答要精确到什么程度。访员追问的一个常见原因是受访者的应答比研究人员希望的更宽泛。明确规定应答的精度可以消除访员追问。例如：

访题：“您认为住在这附近最好的是什么？”

评论：这里必然需要额外的追问，因为这道访题要求获得叙述性应答，但访题又含糊不清，没有说明受访者应答几个要点比较合

适。Groves和Magilavy（1980）发现，与访员效应有关的访员之间最大的差异之一，就是应答这类问题的要点数量。

目前，对产生访员非标准行为的访题有什么特点，并没有一份完整清单。然而，由于访题追问率与访员对结果影响之间有显著关系，因此，改善数据质量的一个有效策略就是使用行为编码对调查工具做好测试（参见第8.5节 ）。在试调查中，研究者能够识别那些可能需要访员追问的访题，并努力改善。减少访员追问和其他行为的需要是最小化访员方差的重要途径之一。

9.5.2　专业化的、任务取向的访员行为

从早期的调查访问开始，研究者就意识到访员与受访者建立和谐的关系非常重要，它能使受访者轻松自在并准确完整地应答。长期以来，人们一直认可这种观点。然而，融洽的关系也有可能是一把双刃剑。一方面，受访者与访员积极的关系会使其希望对调查访谈作出贡献；另一方面，访员与受访者之间的关系在某种程度上因个人性多、专业性少进而导致预设的访谈目标失真，反过来导致调查结果的失真。Van der Zouwen，Dijkstra和Smit（1991）做了一项实验，比较了两组访员的正式的和个性化的访谈，结果具有非常大的不确定性。

我们已经注意到，在有访员参与的情况下，受访者对个人信息或敏感数据如违禁药物使用或健康状况的报告情况通常会比自访问卷的结果糟糕。这表明受访者与访员的关系可能构成报告这类信息的一个障碍，即使是通过电话建立的。

Fowler & Mangione（1990）的研究报告也提供了一些证据，表明当访员的教育地位高于受访者时，二者会形成某种程度的“张力—断裂”性互动，对受访者应答具有建设性意义。然而却没有证据表明在双方地位相当或当受访者地位较高时也会如此。在调查中，融洽问题变得尤为有趣。电访的时间短，访员与受访者之间的互动少，与面访相比，受访者获得访员信息并与之建立关系的可能性明显较小。因此，在电访中，访员与受访者发展融洽关系的潜力更受限制。

总体而言，对什么是融洽关系，怎样才是理想的关系，以及如何影响受访者的表现和数据质量，都缺少系统性的数据。因此，在处理这个问题时，调查机构通常会告诉访员要同时兼具热情和专业性。这一要求的操作含义是：

1）在调查涵盖的主题上，访员应避免发表任何看法或意见。

2）访员应避免提供任何可用于推测与访谈内容相关的个体偏好或价值观念的个人信息。

3）虽然可以通过聊一些非正式的、中性话题来帮助建立沟通，如天气、宠物等，但大部分时间，访员应以任务为中心。

这些准则都相当模糊，而且也没有强有力的证据来证明其对产生标准化数据的重要性。无论如何，调查机构都希望访员可以满足这些原则，以此减少访员对数据的影响。

9.5.3　让访员准确读出访题

提问时统一措辞，也许是标准化访谈最基本和最普遍的原则。有充分的证据表明，访题措辞的微小变化足以影响受访者应答。证明这一点最简单的方法是向对照样本提问相似的、目标一致的、仅在措辞上稍有区别的访题，如：

访题A：在美国，应该允许共产党员在公共场合演讲吗？

访题B：在美国，应该禁止共产党员在公共场合演讲吗？

尽管“禁止”和“不允许”在概念上似乎基本相同，当对照样本被问及这两道访题时，50%的人认为不应该允许共产党员在公共场合演讲，只有20%的人认为应该禁止。这一结果清楚地表明，如果允许访员将访题从“不允许”改写为“禁止”，将会对调查结果产生重要影响（Schuman and Presser，1981）。Rasinski（1989）发现，受访者更愿意支持“援助穷人”的支出而不是“福利”支出，更愿意解决药物“上瘾”而不是药物“康复”问题。虽然这些都是等效的术语，却清楚地唤起了受访者对语句内涵不同的理解。

此外，不管是在小样本还是大样本中，访题措辞的变化也不总是对应答带来很大的影响。譬如，在一项类似的实验中，Schuman和Presser（1981）发现，用“流产”代替“结束妊娠”，效果相同。

另外两类研究则对正确阅读访题的重要性提出了质疑。Groves和Mangione（1980）研究了访员阅读访题和访员相关误差之间的关系，并没有发现二者存在一致的关系。Fowler和Mangione（1990）的报告则说明，高于平均水平的访题误读率与较高的访员相关误差之间并没有联系。

话虽如此，访题措辞的变化会对调查结果产生巨大影响，则有据可查。给予访员改写访题的自由通常会导致较大的访员相关误差。如果访员提问的措辞不同，几乎可以肯定，措辞差异会对应答和结果的造成极大影响。

9.5.4　访员向受访者解释调查过程

培训访员向受访者解释调查过程是另一个减少访员方差的关键技巧，也是理想访谈程序中缺乏理解和具有争议的一个方面。

当研究者对调查中的互动行为进行研究时，他们通常会发现一些由标准化访谈协议导致的尴尬互动（Suchman and Jordan，1990；Houtkoop-Steenstra，2000；Schaeffer，1992）。我们来考虑以下几点：

> 访题：“您如何评价您孩子的学校：极好，很好，好，一般，差？”
>
> 应答：“这取决于您说的是什么意思。我的孩子在读二年级，我非常喜欢她的老师，那是一个非常积极的环境。不过，我真的认为他们并没有做很多数学和阅读练习。另一方面，她非常开心，她喜欢休息和玩乐。”

评论：这里，访员处于两难的境地。受访者对这道访题提出了合理的疑问。访题没有明确指出该评价学校的哪个方面。像许多人一样，对学校、受访者有许多喜欢的，也有许多不喜欢的；受访

者关注的内容将影响到应答。这是访员需要培训受访者的地方，即访员应该向受访者解释调查过程如何进行。这样才可能获得好的应答，比如接下来的对话是——

访员：“这是一个非常好的问题。不过，您无需关注具体事情。在这道访题中，您应该考虑的是，不论您认为访题指的是哪方面，都给出一个最接近您想法的答案。”

访题：“总地来说，您如何评价您的孩子的学校呢：极好，很好，好，一般，差？”

应答：“我认为，不是特别好。”

访员：“我的意思是，您从我读的五个选项中选出一个答案——极好，很好，好，一般，差。我知道这些答案都是人为规定的，可能并不完全符合您的感觉。但是，这个调查的方式就是我们询问人们完全相同的问题，然后，他们从这些一样的答案中选择一个。这样，我们才能对不同人给出的答案作出有意义的比较。如果我们改变了问题，或人们按照自己的方式应答，我们就很难对不同人的说法进行较确切的比较了。”

评论：在创建标准化访谈过程中，这些解释十分重要。访谈是一项不寻常的互动形式，受访者通常很少经历这种互动。事实上，访员询问所有人相同的问题，他们提问的措辞可能与受访者所处的环境不相匹配，可能是武断的，甚至是不合适的。同样，受访者却不得不从这些答案中选择一个作答，可供选择的答案无法像他们自己叙述的那样恰当，此时，似乎良好的沟通被削弱

了（Schaeffer，1992；Houtkoop-Steenstra，2000）。然而，直接向受访者解释这些访题，是让访谈符合标准化的一种方法。

解决此类问题有三种可能的方式。首先，访题清晰是非常有用的。还可以给予访员更多的灵活性，以直接帮助受访者理解和应答。但是，最好的和标准化的方式，则是教访员培训受访者。

9.5.5　访员无指向性追问

当受访者没有完整准确地应答时，访员应该采取一些行动。此时，访员试图获得充分应答的提问被称作“追问 ”（probes）。给访员标准化说明，就是希望访员以非指向性的方式进行追问。追问的原则是不影响受访者选择某个应答的倾向。由于追问是访谈的重要过程，为此，我们提供了以下几个例子进行讨论。

应答类型会对访员面对的追问挑战产生影响。如果是封闭式访题，要求受访者从提供的一组选项中选择应答，则不充分的应答就是受访者并未从提供的选项中进行选择。此时，合适的追问方式是完整地复述访题或应答选项。

如果访题要求的应答是数值，那么，不充分应答就是指受访者未提供应答数值，或给出的应答过于宽泛。此时，追问的目标就是要求受访者用适当的术语应答，或给出更准确的应答。

对于开放式访题，受访者应答不充分有三种可能的形式：

1）没有应答访题，或回答了其他的问题：

访题：“您认为美国目前面临的最大问题是什么？”

应答：“我认为有许多重要问题需要政府要去解决。”

评论：很明显，受访者并没有应答，此时，最好的追问方式也许是重复一遍访题。

这是一个非指向性追问，没有从任何方面改变对受访者的刺激。

2）受访者可能应答了访题，不过，应答太模糊，结果不明确：

访题：“您认为美国目前面临的最大问题是什么？”

应答：“犯罪问题。”

评论：这可能是访题的一个应答，却过于宽泛。犯罪问题可能包括白领犯罪、街头犯罪、高层犯罪等多种情形，研究者希望知道更多受访者的想法。

可能的追问方式：“您能说得更具体一点吗？”或“您指的犯罪到底是什么呢？”

评论：这些追问都是为了刺激受访者阐述和提供更具体的应答。这两种追问方式都没有增加某个应答的可能性。

3）有时访题需要叙述性应答，要求受访者提供几个他们可能有的应答。在上面的例子中，受访者可以应答的数量是开放性的：

访题：“您认为美国目前面临的最大问题是什么？”

评论：当受访者给出“犯罪”应答后，访员可能希望追问是否还有其他问题。这样，一旦受访者给出了第一个问题的完整且可理解的应答，接下来，最好的追问可以是“还有别的吗？”。

4）相比上述非指向性的追问，带有选择项的追问可能更有助于获得应答，不过，却很可能被认为是指向性的。

譬如，对一道有固定选项的访题，就要考虑：

访题：“您怎样评价自己的健康状况：极好，很好，好，一般，差？”

应答：“最近不是非常好。”

指向性追问：那么，哪个选项最接近您的状况呢？一般，还是差？

多数方法论学者将此视为糟糕的追问方式，主要有两点理由：第一，也是最重要的，访员将结论局限在两个差的选项，受访者并没有确切地选择或暗示两者就是其选择。虽然较差的两项似乎与受访者之前的应答比较一致，只是据此将应答局限在两项进行追问，可能不合

适。第二，有证据表明，向受访者再次呈现全部选项会影响受访者对选项意义的理解。例如，若应答仅有3个选择（好、一般、差），几乎可以肯定选择“一般”是最多的，因为与5选项比较，相比原来排在第四更为积极正面。由此可以说，排除前面3个选项后再询问受访者，可能已明显改变了受访者对选项的理解。所以，此时最好的方式是重复一遍访题，至少重复所有选项，以便获得受访者的准确应答。

访题：“在过去的12个月中，您因为自己的健康原因去过几次医生办公室？”

应答：“有几次。”

指向性追问：“您是说4次吗？”

非指向性追问：“您是说少于4次、4次，还是多于4次呢？”

另外，在应答如“美国的主要问题”访题时，考虑下列情形：

应答：“犯罪问题。”

指向性追问：“您是指类似于谋杀和抢劫吗？”

评论：指向性追问意味着增加某个应答的可能性。指向性追问的一个明显标志是以“是”或“否”作出应答。在后两个例子中，指向性追问表现为访员猜测受访者的应答。当访员这样做并提出是否问题时，会导致两种后果。第一，给受访者提供简单的应答方式——不需要受访者再做任何工作，例如，努力回忆他/她

去医生办公室的确切次数。第二，访员可能不经意地表现对某个应答的偏好或确认正确应答。在访谈中，受访者总是试图寻找线索并推测访员想要的应答，指向性追问可能诱导受访者给出访员想要的回答。

已有研究表明，尤其是对运用叙述形式应答的访题，访员需要的追问技巧非常难（Fowler and Mangione，1990）。也有研究表明，访题需要追问的比率越高，受访员影响的可能性越大。从上述例子可以看出，追问是一个极其复杂的过程，有些追问的错误无疑比其他错误更严重。然而，这些证据也充分说明，好的追问和避免指向性追问是访员行为最重要的部分，与访员效应对估计值方差的影响密切相关（Magione，Fowler，and Louis，1992）。

9.5.6　访员忠实地记录应答

完全按照受访者应答进行记录是标准化访谈中访员行为的最后一方面。面对不同类型的访题，访员的任务也不同。Dielman和Couper（1995）对记录误差的研究表明，误差的发生率相当低，用计算机记录比用纸记录的误差率更低（Lepkowsky，Sadosky，and Weiss，1998）。

当访题有固定选项时，访员要指导受访者选择选项应答，同时，不能推测受访者的应答。

访题：“您怎样评价您的健康状况：极好，很好，好，一般，差？”

答案："不错。"

评论：这时访员不应该推测受访者的应答是"好"还是"很好"，而应该重复访题和选项，让受访者从中选择应答。

如果受访者用叙述形式作答，访员则应该尽可能逐字逐句地进行记录。研究表明，当访员只记录大意或结论时，记录差异会很大，并对结果产生影响（Hyman，Feldrnan，and Stember，1954）。

一些调查机构要求访员对叙述性应答进行"现场编码"（field code）。意思是，访员提出开放性问题，没有向受访者提供备选答案，事后访员要对受访者的应答进行归类（参见第10.2.3部分）。有时这些分类非常结构化，并非一项复杂的工作。譬如，一些访题询问的是受访者做某事的次数，应答是一些数字区间。受访者应答为一个确切数值或正好落在某一区间，则访员完全没有必要通过自由裁量来分类，因此是一项非常简单的工作。

另一方面，如果访题结构不是特别清晰，访员要使用的类别就非常复杂，现场编码就特别容易犯错（参见Houtkoop-Steenstra，2000）。有一些机构希望尽可能减少访员对叙述性应答进行分类编码的活动，倾向于让访员逐字逐句记录应答，然后作为一个单独的步骤，由程序员在接受监督的情况下进行编码。另一些机构则对访员的分类行为表现出更大的宽容，访员通常只接受很少的分类编码培训，对其编码活动也几乎没有监督（Martin and Courtenay，1985）。对大多数调查来说，这些都是非常小的问题。然而，一个良好的普遍原则是，减少访员的自由裁量权有利于减少访员误差。因此，尽可能减少访员在编码中的自由裁量权是一个非常好的想法。

9.5.7　小结：减少访员方差的策略

大多数研究者希望访员在面对不同受访者时保持行为一致，且不同访员也遵循统一的行为标准，从而设计了一整套的程序，最大限度地减少访员个人特质对数据收集的影响：逐字阅读访题，无指向性追问，向受访者提供恰当的培训和解释，管理他们的人际互动行为，以及尽可能减少记录答案时的自由裁量行为。然而，这些是否是搜集有效数据的最佳设计，还存在一些争议。

9.6　关于标准化访问的争议

让所有人回答相同的访题的目标，是希望在尽可能一致的条件下用相同的方法进行分析，这也似乎是最大化一致性测量的恰当方式。这是“标准化访谈”（standardized interviewing）的基本原则。这一目标非常看重调查结果的可重复性。“可重复性”（replicability）是科学研究的一项重要特征，允许另一位科学家在独立重复一项研究时，可以用相同的方法得到同样的结果。可重复要求研究者对方法有详尽的描述，以便其他的研究者重复使用。调查研究可重复的关键就在于能高度自信地说明提了什么问题，如何解决问题。

然而，一些批评者也表达了他们对访员行为标准化的担忧：

1）向受访者呈现同样的访题并不意味着向他们表达了相同的意思。事实上，对所有人用一个访题都表达相同的意思，是不

可能实现的理想状态。给访员一些灵活空间，让他们调整访题的表述，以更适合某个受访者，也许能更好地解决这些问题。

2）标准化访谈是一种不寻常的互动，并非常规交谈。最糟糕的是，标准化访谈产生了繁琐且重复的互动形式。如果访员有更多的灵活性，就可以根据具体情形来调整访题，进而更自然地互动。在这种情形下，受访者可能会更自然，并由此产生质量更高的数据。

3）当访员清楚地意识到受访者误解了访题含义时，标准化的约束尤其糟糕。虽然可以设计更好的访题，也是我们追求的目标。不过，访题总是不完美的，受访者的理解能力也不完美。在这些情况下，如果访员及时干预，及时矫正受访者的误解，可能会得到更准确有效的数据。

显然，这三种担忧是有道理的。请考虑下面的例子：

访题："在过去的12个月中，您为自己看病去过几次医生办公室？"

应答："我有两个问题。第一，包括去心理医生的办公室吗？第二，我去过医生的办公室，确切地说，不是去看病，也算吗？"

评论：这两个问题都有道理。对受访者的询问，答案有正确和错误之分。如果最终调查结果有意义，那么，数据使用者就需要知

道，人们在应答时是基于什么作出的判断。

这种互动，是由访题设计造成的，明确指出这一点非常重要。为了获得准确的应答，一道好的访题应该能讲清楚一些要素。

可以说，回答受访者问题的答案就在访题中：如果心理医生是一位执业医师，那么，心理医生的办公室理应被算作医生办公室；第二个问题，去医生办公室，不一定是要看医生。因此，认真读出访题会引导受访者给出一系列具体应答。尽管如此，还是需要说明，在给定访题的情况下，访员应该怎么做。可能的选择有：

1）说“随您怎么理解”。

2）如果访员知道答案，就需要回答受访者的问题，或给出最合适的理解问题的猜测。

“随您怎么理解”对于回答主观陈述类的问题非常合适（尽管当访题包含模糊概念时也不太令人满意），但是，当答案明显存在对错之分时，“随您怎么理解”就显得非常没有说服力。当然，关心访员效应的人担心，如果给了访员自由，访员就可能以各自的方式解释和澄清访题，与受访者自己理解访题比较，就会产生更多潜在的误差。也有人认为，在上面的例子中，大多数访员其实可以帮助受访者正确理解访题，进而得到更好的数据。他们的观点是，给予访员自由产生的净收益一定比访员不一致带来的微小成本大。

事实上，在一些例子中，也有访员被赋予解释访题的灵活性，以此改善访谈程序和数据质量。例如，要搜集某户家庭成员信息，包括年龄、性别、婚姻状况、家庭成员关系等，就不可以明文规定访员在

互动中的行为。同样，调查机构常需要每月从商业部门搜集类似数据。信息是真实的，通常直接从记录文档中导出。在这类调查中，如果需要用到访员，访员和受访者常需要合作填写信息。访员的大部分工作是向受访者说明要求，澄清填写事项和需要用到的判断规则，并且确保所有必填信息都填写完整。访员通常需要逐字读出标准化的脚本。

另一例更适合访员采用灵活访谈方式的是“事件史日历”（event history calendar）。事件史日历是一种显示装置，可以让受访者用来标记对生活（如居住地、教育成就、更换职业、周年纪念日等）有显著影响的事件，这些标记可以帮助受访者回想并注明调查感兴趣事件的日期（Belli，Shay，and Stafford，2001）。使用事件史日历时，访员被赋予大量措辞和追问的自由权。与传统调查提问形式相比，访员也会进行更多的追问。至少有一项研究表明，使用这种方法可以提高数据质量，而且与访员有关的误差也没有增加（Belli et al.，2004）。

上述三个例子有两处共同点：

1）主观性是现实存在，因此，访员应尽量避免通过自由措辞来影响主观陈述类访题的应答。

2）很难预测人们必须提供什么信息，也很难推测以怎样的顺序呈现信息才是合理的。

也有人通过按脚本逐字提问的实验来评估访员澄清的作用，当发现受访者误解访题时，需要通过界定和澄清帮助其理解访题的真实含

义。有证据表明，访员的帮助可以提高某些访题应答的准确性。另一方面，如果让访员澄清访题变成一种常态，就会极大地延长访谈时间，且获得数据质量改善的访题数量也不多。到目前为止，尚没有实例说明在长时的、复杂调查中，如果有大量访题有待澄清，则访员会一以贯之地提供合适的和正确的帮助（Schober and Conrad，1997；Conrad and Schober，2000）。此外，即使是有众多改善数据质量的机会，那么，更简单的方法还是改善问卷访题的设计。

这一争论至今尚未停止。根据Maynard，Houtkoop-Steenstra，Schaeffer和Van der Zouwen（2002）的归纳，把调查访谈视为互动的人认为，标准化访谈概念存在许多内生问题。也可能在一些具体访题上，如果给予访员更多灵活性，帮助受访者澄清和理解访题，既能提高数据的质量，又不会增加与访员有关的误差。同时，似乎研究者也有充分的理由保持谨慎，不允许访员随便改变访题的措辞并过多干预受访者行为，即干预有可能从根本上改变受访者理解访题的方式。大量证据表明，一般而言，访员会在很多界定清晰的任务上表现出不一致性，如追问甚至读出访题的方式。在这种情况下，我们不能乐观地断定访员会恰当地运用这些创新手法。为了成功完成标准化调查访谈，人们争议较少的途径是设计出好访题，由访员按标准化问答过程对受访者进行指导。

9.7　访员管理

前面的部分，我们讨论了访员行为如何影响与访员相关的误差。这一部分，我们将讨论研究者可采取的其他措施，以减少访员误差。具体说来，我们将考虑以下几个方面：

论标准化和会话式访谈技巧

Conrad和Schober（2000）研究了访员澄清访题意涵会对受访者应答产生怎样的影响。

研究设计：20个有经验的访员进行两组不同的电访，每组包括227名受访者。第一组访谈，对所有受访者都使用标准化访谈（例如，逐字逐句读出每一道访题，非指向性地追问）；第二组访谈，将受访者随机分为两部分：其中，一半进行标准化访谈；另一半进行会话式访谈。访员进行会话式访谈时也先逐字读出访题，但可以作出任何有助于受访者理解访题意涵的解释。在20个访员中，有五个进行会话访谈。研究要检验的是，从第一组到第二组访问受访者应答的变化，以及受访者对应答的解释。

研究结果：在第二组访谈中，访谈技巧对应答率没有影响。第一组受访者应答与第二组接受会话式访谈的受访者应答之间的差异（22%）大于其与第二组接受标准化访谈受访者应答之间的差异（11%）。在第一组标准化访谈中，有57%的受访者正确理解了对采购类型的定义，在会话式访谈中有97%的受访者能正确理解，而在第二组标准化访谈中，能正确理解的受访者也是57%。会话式访谈要比标准化访谈多花费80%的时间。

研究局限：由于只有5个访员使用了会话式访谈，因此无法确定研究结果是否适用于有数千名访员参与的研究。这项研究也没有测量应答偏差，仅是两种不同访谈方法的比较而已。

研究意义：相对于严格的标准化访谈，这项研究是对另一种可能访谈方法最好的尝试之一。展示了这些方法潜在的优点以及可能的代价，也说明未来的研究应致力于发现更加灵活的访谈方法。

1）访员的选择。

2）访员的培训。

3）访员的监督与督导。

4）访员的工作量。

5）访员和计算机使用。

9.7.1　访员的选择

为最大限度地优化访员质量，一个明显的方法似乎是选择合适的访员。然而，在近来已知的证据中，少有证据表明选择合适的访员具有重要性。

在美国，访谈通常是一种兼职工作，工资水平特别低，类似于供职于零售公司的销售人员。因此，访员选择的范围受到限制，仅限于劳动力市场上对兼职工作感兴趣的人。

对访员的工作还有一些重要要求。阅读技能和清晰表达能力是访员必备的品质。还有，对于大多数访谈工作而言，由于大多数入户访谈的有效时间在晚上和周末，访员必须保证在这些时间段随时待命，

且愿意一星期工作时间少于40小时。此外，访员在请求受访者配合时要具有说服力，并且需要兼顾调查参与中的各种相关问题。这两项要求也越来越重要。现在的访谈通常运用计算机辅助，故熟悉计算机操作和具备一些打字技巧也非常必要。

除了这些实际工作的要求，并没有足够的证据表明，访员若具备这些特质就必然会有更出色的工作表现。还需要进一步研究证明：改变访员的选择标准是否可以提高搜集数据的质量。毫无疑问，这类研究也必须考虑改变访员选择标准可能要付出的代价。

9.7.2 访员的培训

已有研究表明，访员接受培训的多少对其工作方式的影响非常重要。两个设计相似的实验研究都提供了强有力的证据，表明接受培训较少的访员（少于一天）无法达到让人满意的效果（Billiet & Loosveldt，1988；Fowler & Mangione，1990）。

例如表9.2，此表摘自Fowler和Mangione的研究。在这项研究中，新近招募的访员被随机分配到四组不同的培训方案中，培训分别持续半天、两天、五天和十天。培训项目包括介绍调查目标，访谈提问，当受访者应答较少时应如何追问，应答记录和工作管理总体职责。持续两天及以上的培训都包含模拟访谈，时长十天的培训还会教访员如何使用督导表，如何编码，以及如何有效安排工作日程。

表9.2 访员优秀率或六项标准的达标率与访员受训时长的关系

分级标准	访员的受训时长(天)				
	<1	2	5	10	*p*
逐字读出访题	30%	83%	72%	84%	<0.01
追问封闭式访题	48%	67%	72%	80%	<0.01
追问开放式访题	16%	44%	52%	69%	<0.01
记录封闭式访题	88%	88%	89%	93%	*ns*
记录开放式访题	55%	80%	67%	83%	<0.01
无偏差人际互动	65%	95%	85%	90%	<0.01

数据来源：Fowler和Mangione（1990），第115页。

经过培训，对访员的访谈全程录音，并进行编码，包括如何读出访题，如何追问，如何记录，以及访谈中互动行为的恰当性。从表中可以清楚地看到两个结果。第一，除了在合适的应答选项上做标记以外，在所有其他方面，只接受过半天培训的访员的表现明显不如其他访员，且大多数人读出访题和追问的技巧也不令人满意。第二，我们可以看到，追问，这项对访员来说最难获得的技巧，明显是通过大量培训形成的。

在Billiet和Loosveldt以及Fowler和Mangione的研究中，就访员对数据影响程度而言，分析表明，更高质量的数据和更多的培训密切相关。

9.7.3　访员的监督与督导

对访员的督导同样影响数据质量。在计算机辅助调查中，计算机程序会对没有纳入的数据检查数据的不一致性；尽管可以发现不一致的数据，却很难发现访员在其中的影响。如果是纸版问卷，督导会检查完成的问卷，评估访员是否遵循了培训要求。可在计算机辅助调查中，督导只能检查不同访员的数据缺失率，并以此判断是否存在访员

操作不当。在面访调查中，对于访员在何种程度上严格执行了调查方案，督导员很少能获得持续监督信息。在集中的电访中，督导通常可不时地出现在访问室里。

是否系统地监测问答过程，是督导的关键所在。上述两项研究的一个特点就是，随机对一些访员的访谈过程录音，并由督导进行检查。相反，有些访员的访谈过程既没有录音，也没有督导的监测。很明显，我们会认为，如果没有对访员实施问答的过程进行监测，其表现就不会像被监测的那样好。两项研究都提供了一些证据表明，对访谈录音时，数据质量会提高。这些访谈都是由访员个人独自完成的，对他们来说，录音记录是最实用的监督方法。计算机辅助面访（CAPI）软件最近的一次升级，已经可以随机运用笔记本对访谈互动进行数字化记录（Biemer，Herget，Morton，and Wills，2003）。如果这一技术成为标准化实践，访谈关键部分的行为编码将成为一种常规监管工具。

当然，对在集中的电访中心工作的访员，督导员的监听工作就容易很多。在很多电访中心，让第三方随机听取访谈样本并大体评估部分访谈，是一项标准化的实践。通常，督导员都接受过访谈技巧和调查目的培训，报酬也比访员高。一些监测程序还会自行确定对访谈互动的概率抽样规模，当然也可以由督导来决定。一些监测程序类似于行为编码，可以得出访员和受访者行为的定量数据，也有一些要求督导对访员行为作全面的定性评价。一些监测程序要求督导给予访员反馈，以促使他们在未来访谈进行改善。另一些则会在最后汇总监测报告，提供访员互动结果的反馈。

访员受训时间的长短和督导的多少会影响到调查费用。对访员进行三到四天培训需要的花费比培训时间少于一天的花费会多出三到四

倍。更重要的是，对访员工作进行持续的监测和检查会增加监管的成本。

有可信证据表明，为了让访问员能够以合理的、可接受的方式实施问答程序，就应该保证至少对其进行半天的培训。上面引用的两个研究也表明，对访员表现的偏差进行持续监测，有助于提高数据的质量。然而，在以下三个重要领域却缺乏好的研究：第一，上面引用的两项研究都只针对面访访员，然而，缺少电访访员行为和数据质量的研究。第二，尽管在电访中心做访员监测研究相对容易，却缺乏有效数据来表明监测访员行为对数据质量的影响。第三，据说有证据表明，调查机构提供的培训在数量和种类上存在很大差别，对问答过程的监测也有很大差异。有限的证据表明，培训和督导确实对数据有积极影响，急需进一步研究来证明的是，能起到多少作用，以及监管和培训在多大程度上能改善数据质量。

9.7.4　访员的工作量

为了最大限度地减少访员对方差的影响，对研究者而言，访员的工作量是最后一个可控选择。如前所述，当访员影响受访者应答时，每位访员完成访谈的平均数会直接影响标准误的估计值。即来自访员方差的设计效应，$1+\rho_{int}(m-1)$，在这个公式中，m 是平均工作量。在一项研究中招募更多的访员并减少他们的工作量，是另一个减少访员对标准误估计值影响的方法。

9.7.5　访员和计算机应用

自1980年代中期以来，由计算机辅助进行访问的比例不断增加。访题跃然于计算机屏幕，访员看着屏幕读出访题，并直接在计算机上输入受访者给出的应答。将计算机技术引入到调查过程也引发了一些新的挑战，值得我们额外关注。

第一，我们需要更深入地理解，如何让访员尽可能发挥计算机辅助调查工具的作用。起初，计算机辅助设备只是将纸质工具转置到计算机屏幕上。对访员使用设备的研究表明，一些在纸上运作良好的访题组织方式，在计算机上不仅是低效的，还会产生一些问题。因此，需要进一步研究访员和计算机辅助程序的交互影响。已有的研究表明，访员尚无法使用提供给他们的许多辅助。比如功能键，设计十分有用，访员却从不使用（参见Sperry，Edwands，Dulaney，and Potter，1998）。让计算机更便于访员是研究过程的兴趣所在，为此，需要开发更好的使用规则。

第二，我们需要进一步了解计算机使用会对“受访者—访员”互动产生怎样的影响。我们知道一些显而易见的事实，例如，喜欢使用计算机辅助的访员与受访者的受访经验多少之间并没有消极的相互影响。我们也从对“受访者—访员”互动的观察中得知，计算机成为了参与互动的第三方。访员花更多时间关注的是计算机而不是受访者。我们缺乏研究的是，将计算机这一额外角色引进访谈过程，是否会对受访者行为和数据质量产生重大影响。

第三，毫无疑问，计算机辅助访问技术的诞生对访员培训产生了影响。访员培训的相当一部分精力致力于确保访员熟练使用计算机辅助程序。曾经通过授课、展示和实践进行的培训，现在变成了在计算机上的个别辅导。一方面，计算机可以帮助访员做事，尤其是恰当处理复杂的缺失信息，对于纸笔工具来说这是非常困难的。另一方面，

如果培训时间有限，花时间培训访员使用计算机可能会挤占培训访员掌握其他重要技巧的时间，如争取受访者合作、追问以及与受访者有关的其他事宜。如果没有掌握计算机技能，访员就无法胜任工作。然而，即使没有很好的追问和建立融洽关系的技巧，访员依然可以继续工作。这涉及另一个主题，我们还需要更多的系统信息。

9.8　核查访员的工作

前面讨论的访员管理，旨在减少受访者应答中与访员有关的误差。这一部分讨论的问题可能不是很常见，却非常重要，因为它对数据质量有潜在影响：调查应答或应答搜集方式的访员造假。“访员造假”是指访员未经说明，有意违反既定的访员指南和要求，导致数据污染。“有意”是指访员意识到自己的行为违背了访员指南和要求。

访员造假行为包括：

1）捏造全部或部分访谈——记录由非调查对象提供的数据，并将其作为受访者应答。

2）有意误报处置码和篡改过程数据（例如，记录舍弃的不合格样本；虚报、假报尝试接触受访者的次数）。

3）为回避追问而对访题应答有意进行错误编码。

4）为减少完成访谈花费的精力而有意访问非样本对象或对调查管理机构有意谎报数据搜集过程。

大型调查的访员多是临时招募的兼职人员，他们不会参与后续研究，因此，很可能造假。很少有造假的研究发表，表9.3显示了美国人口普查局的一项研究（Schreiner，Pennie，and Newbrough，1988）。在这项研究中，造假包括访谈未经抽样的对象，或利用研究未认可的方式进行访谈。表中显示，与不断进行的当前人口调查（CPS）和全国刑事犯罪受害调查（NCVS）相比，住房闲置调查属于一次性数据搜集方式，可能有更高的造假率。

表9.3　美国人口普查局三项调查中的访员造假比例

调　查	造假比例
人口现状调查	0.4%
全国刑事犯罪受害调查	0.4%
纽约市住房闲置调查	6.5%

Schreiner，Pennie和Newbrough（1988）也发现，没有经验的访员比有经验的访员更易造假，有经验的访员则有更精密的欺骗模式（例如，伪造追踪调查的第一轮数据）。

我们似乎可以通过多种管理程序来减少访员造假。减少访员造假最常用的方法是对访员进行培训，强调数据搜集协议中诚实的重要性。培训之后，再通过数据搜集过程中的核查程序予以补充。有三种主要的审查方法：

1）观察法。

2）重访法。

3）数据分析法。

“观察法”指在访谈过程中有第三方（除访员和受访者之外）在场监听或观看。最常见的观察法是由督导员在电访控制中心随机抽取一个访谈子样本进行监听。比较常见的是监听5%～10%的访谈。这种观察技术是惯常做法，并且被认为可以有效遏制电访中的访员造假行为。在面访中，目前，一些计算机辅助面访使用电脑内置的数字化记录设施记录随机选中的访谈（Biemer，Herget，Morton，and Willis，2003）。然后，由调查管理部门检查这些录音并尽力发现有潜在造假迹象的反常表现（例如，笑声和有人在一旁评论表明访员可能选择了朋友作为受访对象）。

“重访法”是指为了检查访员是否按规定完成访谈，在访谈结束后，由另一位工作人员（通常是督导员）对受访者进行重访。这是分散访谈设计最常用的技巧，不过，因对应答率要求较高也面临着挑战。面访的重访十分昂贵，尽管会产生最高的应答率；电访重访相对便宜，应答率却较低；邮寄自访问卷的回收率最低，成本也低廉，可以适用于更大比例的待检查对象。在审核中，有时会出于成本的考虑而选择使用某些混合方法。

“数据分析法”是指检查访谈获得的完整数据。有时也检查过程数据。例如，如果一个访员执行访谈的时间非常短，如果第一次联系受访者就同意接受访谈的比例非常高，或者如果每次访谈时长都低于平均水平，那么这个访员就可能要重点关注和后续审核。有时，也会审核调查应答的模式。对非常规应答模式（如关键分叉式访题的应答会导致问卷的大幅度跳转）也要审核。

美国统计协会和美国舆论研究协会（AAPOR）提供了审核访员造假的实践指导手册（参见http://www.amstat.org/sections/SRMS/），包括了部分以下建议：

1）应确定一个造假审核的概率样本（如5%～10%的样本）。

2）审核问卷应包含住户结构和/或其他资格要件；数据采集模式；访谈时长；激励报酬（前提是有报酬）；数据采集中的计算机使用；讨论要点；关键项目，尤其是控制访谈大幅跳转的访题。

3）如果某个访员在抽样审核中没有通过，则要对其全部工作进行审核；在审查期间，该访员不能参与数据搜集活动。

4）如果大多证据都指向访员造假，应该在组织规则允许范围内立即实施访员的人事调整。

5）修正所有被发现的造假数据。修正的方法通常是进行完整访谈。

6）调查的技术报告应该有概率样本获得的造假数据的比例。

造假报告是标准的调查文件的一部分，调查专业组织希望借此增强人们对这一问题的意识，并找出解决办法。

对访员造假，还有很多未回答的问题。由于只是对已完成的访谈样本进行审核，因此，实际发现的造假比例与审核的抽样比例有关。人们普遍相信，任何访员都不会总是造假，而工作量的差异又使得抽样变得十分困难。最佳的抽样比例是什么？是否可以用适用的抽样过程替代？访员欺骗的动机是什么？奖励机制是否产生了更大的欺骗（例如按已完成的访问支付访员固定报酬可能会比按小时支付更鼓励造假）？对这些问题，都需要更好的解释。然而，此时我们想说，审核访员是如何执行任务的，应该作为访员管理的一个常规部分。

9.9 数据搜集中录音资料的使用

本章最后一节介绍模拟或虚拟访谈。传统上，数据搜集通常有两种类型：要么访员读出访题并记录应答，要么受访者阅读访题并写出应答。然而，有一些数据采集方案融合了这两种方法。通常，方案包括了事先的录音（或可能是计算机声音），向受访者大声读出访题。有时候，受访者使用计算机键盘作答；有时候，则使用按键电话作答；又有时候，用一种声音识别系统对应答解码。这些数据搜集方案的共同特点是，受访者在没有访员帮助的情况下作答。为了进一步消除不同数据搜集方式的界线，如果受访者盯着电脑屏幕，他们将看到事先编写好的程序语言，并随之看到屏幕上呈现出一张人脸图像。从图像中可以识别访员的性别、种族和年龄等特征。尽管从声音就可以推断出这些特征，有了图像会更加明确。

这些方案真的属于访谈这一章吗？这一节的存在说明，我们认为它是值得讨论的，虽然讨论比较简短。

作为自访问卷的辅助工具，计算机访谈声音可以帮助阅读能力受限的受访者理解访题。计算机声音能够使用多种语言流利地阅读访题，这在访员在场的访谈中是很难实现的。还能标准化访谈过程管理的各个方面，这是现场访员不可能控制的，如发音、节奏和语调。同时，也很容易最大限度地减少无意识的闲聊，这些闲聊都是潜在偏差。如果受访者需要帮助，也可以提供帮助，如术语定义，并能让所有受访者得到的定义一致。如果增加访员图像，访员的年龄、性别和种族也是可控的。

对一些调查而言，这些可能都是加分项，也有很好的例子说明计算机辅助自访问卷解决了访员作为数据搜集者可能的局限。然而，对大多数调查而言，这些仅仅是辅助自访问卷的策略。受访者对计算机声音的反应是，好像自己也是计算机，而不是活生生的人。如前所述，当要求受访者提供可能尴尬的信息时，计算机辅助具有优势。受访者也不会特别关注计算机上的访员图像。受访者面对计算机提供的“女性”，并不会产生与面对女性访员相同的性别影响。

然而，计算机声音并不擅长建立合作关系。相比与现场访员交谈，受访者更倾向于中断与计算机的访谈。计算机声音能够通过编程可靠地安排提问次序，却不能解决其他个性化问题。尽管没有很好的证据，但根据合作率与访谈中断数据可以假设，计算机并不如现场访员善于激励受访者。

进一步的研究应该是，如何更好地融合多种数据搜集方法；让计算机在某些方面能与访员相似，也可以作为研究议程的一部分。然而，可以确定的是，在未来很长一段时间内，访员还是调查过程的重要组成部分。如何筛选、培训和管理访员以保证更好的数据质量，依然是调查方法重要的议题。

9.10 小结

访员承担着实施许多调查设计特征的责任，这些特征会对覆盖率、无应答和测量误差产生影响。一些测量误差有系统原因，另一些则来源于不同访员的差异性。系统误差类似于少报社会非期许行为，多报基于访员外表特征（如种族、年龄和性别）判断的符合访员预期的应答，更多的测量误差则来自于有经验的访员。

当访员行为对应答影响不可控且多变时，调查结果中与访员有关的方差会增加。对不同访题，访员行为的重要性不同，一些访题比另一些显示出更高的访员方差。除非进行专门测量，否则，这些来自访员可变效应对数据的影响是不可计量的。

访员方差提高了调查统计中的标准误，进而使调查统计的准确性降低。在访员工作量很大时，访员误差对标准误的影响尤其值得注意。有些研究靠极少量的访员（极端情况下只有一个）完成所有访谈。这种情况下，存在严重访员效应的可能性极大。与此相似，在大规模调查中，每名访员进行50、100甚至200次访谈的情形并不罕见，此时，访员对标准误的影响也可能很大。

访员有效地执行下列标准，可以减少来自访员的统计误差：激励受访者的良好表现，互动中采取任务导向法，逐字逐句读出问题，向受访者澄清访题，无指向性追问，准确记录应答。有争议的是，标准化访员行为是不是一定可以最小化误差。这些争议表明，还需要进一步的研究来探讨访员对访谈过程的影响。

近年来，研究者在访题设计上再次展现出兴趣。然而，在搜集高质量数据过程中，仍然有忽视访员角色的倾向。随着调查越来越多地依靠电访，人际因素的影响大幅减少，随着越来越多地使用计算机辅助搜集数据，我们有理由认为，访员的作用也变得越来越不重要。然而，无论采用哪种数据搜集方式，受访者的动机和对角色的认知都会影响到他们的行为。此时，访员就会在激励和引导受访者方面展示其重要性。

关键词

访员方差（interviewer variance）

追问（probes）

可重复性（replicability）

交叉样本分配（interpenetrated sample assignments）

标准化访谈（standardized interviewing）

事件史日历（event history calendar）

进一步阅读资料

Conrad, F., and Schober, M. (2008), *Envisioning the Survey Interview of the Future*, New York: Wiley.

Fowler, F. J., and Mangione, T. W. (1990), *Standardized Survey Interviewing: Minimizing Interviewer Related Error*, Newbury Park, CA: Sage.

Maynard, D., Houtkoop, H., Schaeffer, N., and van der Zouwen, J. (2002), *Standardization and Tacit Knowledge*: Interaction and Practice in the Survey Interview, New York: Wiley.

作业

---

1. 40位访员完成了2 000次访谈。某特定选项的ρ_{int}值，即与访员有关的组内相关系数为0.015。

   （a）在访员平均工作量的前提下，计算该项目的访员设计效应（DEFF）的值。

   （b）简述设计效应的含义。

2. 定义和描述4项培训访员标准化其行为的基本技术。

3. 通过行为编码，可察觉出哪种行为显著地与应答的访员效应关系最紧密？

4. 针对下列访题的每个应答，写出您认为标准化的、无指向性的访员提问方式。

   访题：您认为美国目前面临的最大问题是什么？

   应答：塔利班。

   应答：不知道。

   应答：毫无疑问，犯罪和恐怖主义。

   应答：考虑到国外正在发生的事，我认为不必担心自己的问题，而必须投入所有力量去应对恐怖主义。

5. 您是一项大型家庭调查项目的管理者，调查通过面访测量您国家的贫困水平。您已经认真制订了一份交叉配置方案，将访员随机分配到初级抽样单位家庭样本子集。每位访员平均完成25个访谈。

在这项设计中，您已经计算了反映访员方差的ρ_{int}（根据本章的模型）。

下表提供了一些结果：

变量	ρ_{int}
贫穷家庭百分比	0.09
有失业成年人的家庭百分比	0.02

项目团队收到了一个坏消息：明年的预算可能削减。请您评估，当访员数量减半而样本规模不变时，对估计值精度的影响。

6. 一项调查旨在测量家庭样本中成年男性和女性的生育经验和性行为。其中，讨论了需要最小化与访员性别有关的测量误差。在您决定全都使用男性或女性访员，还是男女访员数量随机混合时，需要考虑什么问题？

7. 哪种数据搜集方法对访员欺骗行为作用不大：面访还是在电访中心进行电访？（请给出理由）

8. 您是一项大型调查的管理者，正面临一个问题，即是否要由面访改为电访。您还参与一项两种方法比较的经验研究。一半的样本（大约1 000个）使用电访，另一半则用面访。电访中心使用25名访员，面访团队拥有50名访员。电访中心不间断地监测访员并在必要时启用补救培训方案。面访则采用全样本核查，并选取每位访员的访谈进行一次观察。针对两种方案，您已经制订了一套访员交叉配置方案，以测量访员方差。

调查的最后，您对关键估计量计算了访员的组内相关，发现电访组样本显示的平均组内相关为0.025，面访组为0.040。如果只使用访员方差，您更偏向哪种访谈方法，为什么？

9. 您正在检查您指导研究的访谈行为编码数据。您关注以下访题："最近12个月，从2008年4月20日到现在，您与医护人员就自己的健康状况有过多少次交谈？包括医生、护士和其他医疗人员；包括电话交谈和当面交谈；包括健康问题和积极的健康经验。"

   您发现，对这道访题，访员逐字读出访题的比例只占70%，比其他访题逐字读出的比例更低。您会：

   （a）什么都不做。

   （b）为提高标准化提问比例，提供补救性培训。

   （c）尝试修改访题。

   评价上述每一种可能的行为。

10. 就您所学，对访员培训时间应更长更密集的观点进行辩论，并说明长处和短处。

11. 为最小化访员对调查估计值的影响，研究者应该做哪四件事？

12. 虽然标准化访谈被认为是调查研究的"最佳实践"，也有人持反对意见。简述双方的主要观点。

13. 就您所学：

    （a）定义"访员方差"。

（b）找出一种方法，研究者可以评估访员方差对测量误差的影响。

（c）简述访员方差如何影响调查估计值的准确度。

14. 简述（用2～3句话）下列设计特征是否影响访员效应的大小：

（a）问题的形式。

（b）访员的特征。

（c）计算机辅助访问。

15. 为什么集中化访谈可能减少也可能增加（统计量）总方差中的访员影响，请分别给出理由。

16. 针对下列访题的每一个应答，您认为一个合格的访员接下来应该说什么：

访题：总体上，您如何评价自己目前的健康状况：极好、很好、好、一般、差？

应答：不是很好。

应答：嗯，一般而言，我认为自己的身体相当不错，但是今天我感冒了。

应答：我没有以前跑得快了。您是这个意思吗？

应答：嗯，我的确有糖尿病，但除了这一点，我要说，很健康。

应答：像大多数我这个年纪的人一样，我有病痛，总体上，我想说，很健康。

17. 如何定义“无指向性”追问？

18. 如何定义“指向性”追问？

10　调查数据的后处理程序

10.1　导言

本章讨论数据搜集之后的那些事儿。这些步骤的最终目的是估计目标总体的属性。在早期调查中，要运用大量的工作人员核查每一份问卷与应答，不过，随着时间的推移，人们逐步形成了一些程序，减少了这些步骤的负担。

说明实地调查结束后工作步骤的一个途径，是把纸版问卷调查与某种计算机辅助调查进行比照。图10.1展示了纸版问卷调查的一般流程。纸版调查是用纸作为工具记录受访者应答。由于调查的产出是数字，如果数据不是数字形态（如开放问卷的文字应答，应答措辞的选项框），还必须将其转换为数字形态（这个操作即“编码”［coding］）。在所有数据都表现为数字形态以后，还需将其通过输入变成电子文档形态（即图10.1中的“数据录入”［data entry］）。

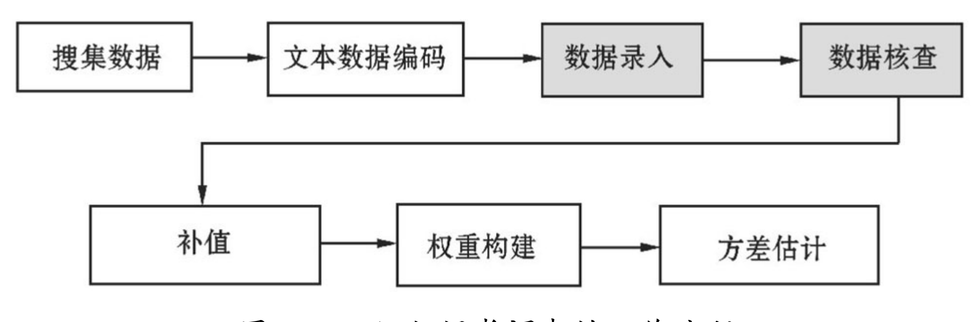


图10.1　纸版问卷调查的工作流程

即使数据已经转化为数字形态，也会有一些基本约定（如1=是，2=否，8=不知道，9=拒答）。有些应答与其他应答之间应该有逻辑关系（如应答为“男性”的就不应该报告剖宫手术）（即图10.1的“数据核查”［edit check］）。如果有访题数据缺失，研究者可能要用一个估计值补值（即“补值”［imputation］）。有些设计在对目标总体进行统计估计时还需要进行数据加权，这项活动也在数据搜集完成之后（即“权重构建”［weight construction］）。为了评估调查估值（即“方差估计”）的质量，在数据搜集完成之后就要进行调查统计精度的初步估计。尽管有些操作在部分问卷返回之后就已经开始，不过，一般而言，操作的完成还是要等到问卷全部返回、数据搜集活动完成之后。

如果运用计算机辅助调查（CATI，CAPI，Web），则操作过程有明显的变化，有些工作直接纳入了数据搜集过程中。图10.2显示，在纸版问卷中，数据核查是数据搜集完成之后的一个步骤，在计算机辅助调查中，则是与数据搜集过程同步完成的。注意，在图10.1中，数据录入与核查在数据搜集之后，计算机辅助把传统数据搜集之后的活动前置，让调查更加聚焦于数据搜集。值得注意的是，即使在计算机辅助调查中，研究者还是会在数据搜集之后进行某种数据核查，参见第10.4节 。

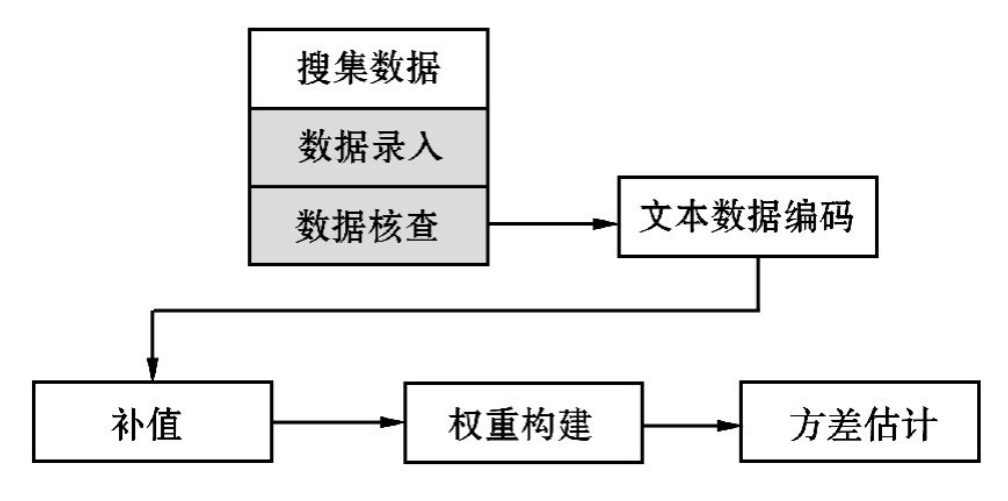


图10.2　计算机辅助问卷调查的工作流程

计算机辅助调查对操作顺序的重置，让涉及数据录入和核查的决策在问卷开发阶段就得做出。数据搜集、录入、核查的合并还意味着调查搜集的只有数字化数据（即没有文本数据）才会跳过编码阶段。由于在调查机构中，编码员属于另一个部门，为此，人们更倾向于避免使用开放访题以及其他需要文本应答的访题。进而，越来越多地调查只搜集数字化的数据。最多也就在封闭访题中运用“其他，请注明”之类，以让受访者提供不在选项中的应答。

这一章讨论在数据搜集之后可能要做的6项活动：

1）编码——把文本应答转化为数字应答的过程。

2）数据录入——把数字转化为电子文档的过程。

3）核查——检查记录的应答，查找错误和不一致。

4）补值——通过给出应答值来填补访题缺失值。

5）加权——通过调整调查统计值的计算来减少因无覆盖、无应答、非概率选择样本带来的损害。

6）抽样方差估计——对计算调查统计值（依据设计，任意可测量的统计误差）的不稳定性进行计算估计。

10.2　编码

编码 （coding）就是把非数字数据转化为数字数据。这一节将讨论调查方法中的编码，作为一种实践，编码又如何影响了数据质量与调查成本。有时候，把非数字数据转化为数字数据非常简单，譬如在自访纸版问卷中：

请说明您现在的状态（在合适的选项上打勾，单选题）：

□全日制学生。

□非全日制学生。

□申请者，已拿到录取通知。

□申请者，未拿到录取通知。

这里，编码就是给4种可能应答的每一种赋予一个足以相互区别的代码（如1=全日制；2=非全日制；3=已拿到录取通知书；4=未拿到录取通知书）。这些数字在电子文档中就变成了这些选项的值。

有些时候，转换会相对困难。譬如，全国刑事犯罪受害者调查（NCVS）的访题：

事件发生时，您工作所在的公司/政府部门/商业组织/非政府组织的名字是什么？

如果是商业组织，属于哪个行业？

如果有必要，请读出：他们生产什么，事件发生时您在什么位置工作？

您做什么工作？即事件发生时，您在从事什么工作？

譬如：管道工，打字员，农场主。

在这份工作中，您通常做什么？

访员要把受访者对这些访题的应答写下来。譬如，受访者对访题“如果是商业组织，属于哪个行业？”的应答是：“为不同尺寸的容器生产塑料盖。塑料盖用来封住容器内的液体，使其不至于外泄。有时，容器会用来装有毒液体，因此也可能给某人带来伤害。”在自访问卷设计中，文字是受访者自己手写或输入的。在这种情况下，会有一部分人的应答不在分类中，进而影响统计值的计算。譬如，在上述应答中，有一类是“塑料和树脂制造”，也许与应答符合。在北美工业分类系统中（North American Industry Classification System，NAICS），这一类的编码是325211。调查中，也许会计算受访者中应答325211的比例。（更可能的是计算更高一个分类层级的累计数。）

文本编码对调查结果的统计分析能力至关重要，而编码本身也会产生误差，并因此对调查统计值产生可见的影响。例如，Collins（1975）发现，在连续月度访谈中，32%的受访者经历了职业编码（运用3位数职业编码）的变化。其他以月计的职业变化显示，许多这类差异性是不同月份的编码差异性导致的。

非常重要的是，在调查中，不是所有非数字数据都是文本，也有可能是图像（如照片、录像）、声音（如磁带）或需要数字化描述实物样本（如土壤样本、血样）。

10.2.1　编码的实践问题

编码既是翻译，也是归纳。正如所有的翻译一样，需要把一个框架下的实体映射到另一个框架下。如果两个框架兼容，则映射比较容易。如果两个框架不匹配，则翻译就变得复杂，也一定会有误差。编码式的归纳，就是单个的应答综合为一个编码类。正如所有的归纳一样，人们必须确定在哪个层次上综合才适合运用编码。

因此，第一个需要关注的就是对文本数据进行归类的分类框架的构建。不幸的是，对这个框架，还没有一个大家都能接受的术语。有些人把它叫做编码结构（code structure），有的叫做“命名”（nomenclature），还有的叫做“编码列表”（code list）。有用的是，编码必须具有下列属性：

1）一个唯一的数值，用于稍后的统计计算。

2）一个文本标签，用于描述类别性的所有应答。

3）能穷尽所有应答（必须把所有应答都纳入某一类）。

4）互斥性（任何一个应答都不能被纳入一个以上的类别）。

5）符合分析目的的分类类别。

鉴于调查的科学目的（如果研究的目的是发现因果过程），每一个编码类都要与关键假设的不同部分有关联。例如，要研究监督的成就，则在职业编码中就需要把监督者和非监督者进行区分。另一方面，如果研究问题涉及职业分类相关的教育背景，则需要依据教育背景进行职业编码。

用于编码的分类数也需要明确。极端的情形是有多少应答者就有多少类，不过，这种方法通常会让事情变得复杂，因为没有归纳。因此，类别的数量通常因变量的运用而定。简单地说，编码系统的构造是实质性的重要活动。对数据的不同运用需要不同的编码结构。

对同一种测量可以有多种编码结构。例如在英国，职业访题的应答通常会有两个编码。一个是标准职业分类编码，另一个是社会经济群体分类编码（地位排序，也依据雇主的规模）。从访题获得的多种应答，还需要针对每一个用到的应答创建单独的编码变量（例如，“美国目前面临的最主要问题是什么？”）。

无论用于分析的编码结构有多复杂，总会有一个应答纳入不到编码中。为了减少不能纳入编码应答的数量，通常会从先期完成的问卷中提取一组应答建构编码结构，进行测试、改进，这样会逐步成熟。如果有应答不能很好地纳入到分类中，通常就要对既有的编码结构重

新考量。在编码过程中，如果编码结构发生了改变，那么研究者就不得不对之前的编码进行检查。

最后，非常重要的是，编码结果应对的是所有应答，即便是不规范的应答。对某些受访者的无应答也要有一个编码（“不确定”）。此外，如果问卷允许受访者依据对之前访题的应答判断某道访题不适用而跳过（如受访者应答不曾遭受过盗窃，则与盗窃相关的访题就可以跳过），那么，非常重要的是要有编码识别访题不适用的受访者，进而判断数据文件中的无应答是正常的（“不适用”）。由于这一选项涉及大量变量，因此最好用一个统一的数值编码。例如，如果是一位数编码，通常会用“9”编码“不确定”，用“0”编码“不适用”。如果是两位数编码，通常会用“99”编码“不确定”，用“00”编码“不适用”。此外，对某些变量，如果因为特殊原因没有获得应答，调查者也许希望把它挑出来。譬如受访者拒答或者说不知道。在这种情况下，同样，也需要有单独的专门编码。

10.2.2　编码活动的理论问题

编码的基础就是一个决策，即确认两个表达是不是等价的。例如，“我修复排水，安装水管，更换水池”与职业便签“管道工”是不是等价的。在理想的情况下，编码员的分类应该与受访者提供的分类完全一致。图10.3显示了问题的特征，如果编码员的分类选择（右下角方块）与受访者的预计的分类（上方的方块）完全一致，则编码具有准确性。实际的挑战是，编码员看不到受访者的预期，只看得到访员记下的文字（在自访问卷调查中，则为受访者记下的文字）。

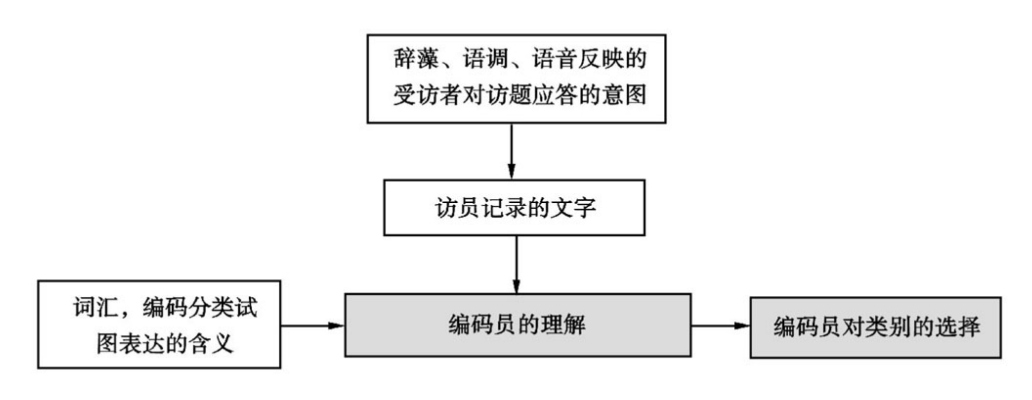


图10.3　编码员的理解与判断

编码，就是编码员试图识别受访者通过记录的文字意图。这还只是挑战的一半，因为编码员的任务是把受访者的意图映射到一个（仅一个）编码类别。要做到这一点，编码员还需要完全理解研究者界定每一个编码类别（左下方块）的意图。这个理解，是编码类别描述的质量和培训编码员质量的函数。

10.2.3　“实地编码”——中间体设计

正如我们在7.2.4节和9.5.5节学过的那样，开放题让受访者使用自己的语汇应答。有时，正揭示了受访者对访题的理解或受访者的记忆结构，这正是对研究者的研究有价值的素材。封闭访题向受访者展示了研究者试图发现的应答，这些应答也许适合，也许不适合受访者。

实地编码 （field coding）指受访者应答开放访题、访员将应答转化为数字类别的过程。如此，在实地编码中，既要搜集应答，还要编码。例如，全国刑事犯罪受害者调查询问一道开放访题，受害发生在什么地方，在获得应答之后，访员就得直接选择合适的编码（参见表10.1）。

表10.1　NCVS调查中，访题“这次伤害发生在哪儿？”的实地编码举例

分类号	分类定义
在受访者家里或住处	
01	在受访者自有的住处、车库、封闭阳台(包括非法进入或试图非法进入)。
02	在受访者自有的另外建筑,如单独的车库、储藏室等(包括非法进入或试图非法进入)。
03	在度假屋/第二住处(包括非法进入或试图非法进入)。
04	住在酒店或汽车旅馆的房间(包括非法进入或试图非法进入)。
靠近自己的家	
05	自己的院子、人行道、车道、停车棚、未封闭的阳台(不包括单元住宅的院子)。
06	单元住宅的大厅、储藏区、洗衣房(不包括单元住宅的停车位/车库)。
07	在紧邻自家的街道。
在或靠近朋友/亲属/邻居的家	
8	在他们的家或另外建筑。
9	院子、人行道、停车棚(不包括单元住宅的院子)。
10	单元住宅的大厅、储藏区、洗衣房(不包括单元住宅的停车位/车库)。
11	在紧邻他们家的街道。
商业空间	
12	在餐馆、酒吧内。
13	在银行。
14	在加油站。
15	在其他商业空间,如商店。
16	在办公室。
17	在工厂或仓库。
停车场/车库	
18	商业性的停车场/车库。
19	非商业性的停车场/车库。
20	单元房/联排房的停车场/车库。
学校	
21	在学校建筑内。
22	在学校区域内(学校停车区、玩耍区、校车等)。
开放空间,街上或公共交通	
23	单元住宅的院子、停车场、场地、玩耍区(非学校的)。
24	街上(非紧邻自家/朋友/亲属/邻居的家)。
25	公共交通或公共交通站(汽车、火车、飞机、机场、航空站)。
其他	
26	其他。
27	剩余的。
28	超出一般的。

由于受访者可以用任意有用的表述应答，访员就得说明访题、评估可能适合的类别，在不确定时进行追问，然后在26个不同应答类别中选择一类。这显然不是一件轻松的事。另一方面，办公室编码员面

对的只是访员记下的文字，而实地编码则可以通过追问获得帮助，且听到的应答更丰富。

Collins和Courtney论实地与办公室编码

Collins和Courtney（1985）比较了搜集数据时进行实地编码和数据搜集完成后进行办公室编码的结果。

研究设计：在调查之前，把7道访题随机分派给3个重复样本：（1）开放题实地编码；（2）开放题办公室编码；（3）封闭访题。观察3个样本中实地与办公室编码混合的差异，把6个面访访员随机分派到4个样本区域。比较180～200人重复的应答分布和访员方差。

研究发现：如果办公室编码员和实地编码员使用相同的编码分类，两者之间没有系统误差。有差别的是，办公室编码员的某些编码使用率略高。每个受访者的编码平均值相似，不过实地编码使用“其他”类的略高。在应答总方差中，实地编码的访员差异性贡献大约为3%（$deff_{int}$ =1.5，工作量为25），而办公室编码的仅有0.6%。

研究局限：办公室编码框仅限于问卷提供的分类，或许对办公室编码员有更多限制。由于没有测量办公室编码员的方差，便只能比较$deff_{int}$。没有讨论访员的实地编码培训。

研究意义：研究指出了实地编码的局限，即在访谈中，一边要听应答，一边还要从一组固定的分类中选择应答所属的类别，进而

加重了访员的工作量。

与基于书面或磁带记录的编码比较，少量的研究涉及了访员实地编码的过程。这些研究比较了完成相同工作的两种方法，以及两种方法结果的一致性。例如，对职业编码而言，办公室编码员的编码与实地编码比较，更接近于金标准（职业编码结构中，有近400个类别），不过，两者之间的差异不大（Campanelli，Thomson，Moon，and Staples，1997）。尽管如此，也有证据表明实地编码对访员行为有负面影响（参见Maynard，Houtkoop-Steenstra，Schaeffer，and van der Zouwen，2002；Fowler，and Mangione，1990）。所有造成测量误差的访员方差因素（参见第9.3节 ）也都适用于编码。也就是说，由于一个人承担访员和编码员两方面的工作量，如此，在调查估计中，编码活动会造成额外的不稳定性（即“编码员方差”）。正因为如此，办公室编码和实地编码就变成了成本与分类误差之间的一个平衡。

10.2.4　标准分类系统

这一节讨论两个常用的分类系统，一个是职业测量，另一个是行业测量。在社会和经济调查中，经常要测量这两个变量，国际机构也制定了分类标准并定期更新。标准化非常有价值，便于对一般人群的各种调查进行比较。

标准职业分类（The Standard Occupational Classification，SOC）。在美国，劳工部负责管理标准职业分类系统。在1980年有过一

次修订，然后是1994年、1999年，反映的是经济组织中职位的变化。运用职业分类，可以依据研究目的把劳动人口划分为不同的职业组以及管理记录系统。其编码系统包含有820个职业，合并为23个大类，98个小类，以及449个一般职业。每个一般职业包括了详细的职位要求、技能、教育程度、经验。表10.2列出了标准职业分类的23个大类。

表10.2　23组标准职业分类

分类号	分类定义
11-0000	管理职位
13-0000	商业与投资操作职位
15-0000	计算机和数学职位
17-0000	建筑与工程职位
19-0000	生命、物理和社会科学职位
21-0000	社区与社会服务职位
23-0000	法律职位
25-0000	教育、培训和图书馆职位
27-0000	艺术、设计、娱乐、体育、传媒职位
29-0000	健康照料实践与技术职位
31-0000	健康照料支持职位
33-0000	预防服务职位
35-0000	食物准备与服务相关职位
37-0000	建筑、地面清洁与维护职位
39-0000	特护与服务职位
41-0000	市场销售及相关职位
43-0000	办公室与行政支持职位
45-0000	农业、渔业及林业职位
47-0000	构建与拆除职位
49-0000	安装、维护以及修理职位
51-0000	生产职位
53-0000	运输与物质移动职位
55-0000	军事—具体职位

最新的SOC版本采用了4个不同的组织原则。第一，也是1980年以后SOC的基本概念，依据实际从事的工作。第二，吸收国际标准职业分类，体现工作的全球化，把“妇女工作领域”进行了区分。第三，设计了一个“技能系统”，区分每项工作运用的不同技能。第四，运用宏观经济理论，识别一个“经济系统”。新系统有以下用户：教育和培训计划师、猎头、学生以及其他职业搜寻指南者；各种政府计划；以及希望重新设置薪酬的私人公司。目标还是继续关注实际从事的工作类型，且相信，作为一种组织原则，基于技能的职位评估并不十分准确。同时，也做了很大的努力试图反映自1980年以后出现的职业类型（譬如环境工程师、护肤专家、门房、健美指导）。

用于构建编码的编码计划背后，有各种假设（参见：http://www.bls.gov/soc/socguide.htm）：

1）分类要覆盖所有的有酬工作，包括在家族企业中非直接付酬的工作，不包括志愿者职位。分类的最后一层，每一种职业就是一个职位。

2）依据实际从事的工作、技能、受教育程度、训练以及证书分类。

3）专业和技术工人的监督者通常与其监督对象有着相似的背景，故两者是一类。同样，班组长花至少20%的时间带领的生产工人、监督者、销售人员、服务人员与其是一类。

4）生产、服务和销售一线的管理者和监督者要花80%的时间从事监督活动，却要分别划为合适的监督者类别，因为他们的监

督活动与其监督的工人有很大区别。一线管理者是发挥监督和管理作用的小机构管理者，如财务、销售和人事工作者。

5）学徒和学员可划为正在受训的职业，不过，帮助其学习的人则要单独分类。

6）如果一个职位的职责在结构中不明确，则可归于合适的剩余类别。

7）如果公认可以被归入多于一个类别，则应归入对技能要求最高的一类。如果对技能要求没有可测量差异，则应归入其花费时间最多的类别。

8）数据搜集与报告者对职业的分类越细越好。不同的人或许会依据其搜集数据的能力和使用者的要求运用不同层级的累计。

这些评论强调，分类的目的就是依据劳动人口在工作中运用相似技能和发挥相似职能的差异将其归入到不同的类别。在SOC中，需要注意的是在不同的组织中，职业名称或与工作名称不同。例如，“生产助理”在电视台与钢厂也许有着非常不同的职责。的确，在构建职业分类中始终存在的一种难题是编码结构在多大程度上反映了工作的环境（如一个行业）、技能特征，以及实际发挥的作用。这就是说，编码结构应该反映职业特征。

北美行业分类系统（The North American Industrial Classification System，NAIC）。与标准职业分类系统一样，传统上，行业分类标准化是政府统计部门做的。随着经济的全球化，为了比较不同国家的宏观经济结构，需要统一行业定义。影响美国统计的

最新版本是1997年实施的北美行业分类系统。这是与联合国制定的国际行业分类系统（the International Standard Industrial Classification System，ISIC，第3版）相对应的分类系统，包括了加拿大、墨西哥、美国的共同努力。

正如每个编码结构的改变一样，修订的动机来自于既有测量概念与可用编码结构的不匹配：

1）既有分类缺少许多新商业类型，不能反映美国的经济。

2）实施北美自由贸易协定需要比较加拿大、墨西哥和美国的行业生产。

北美行业分类系统包含1 170个行业，其中565个属于服务部门。分类系统为一个6位数编码系统，其中的5位数在3个国家间是可比较的。头两位数为最高层级的累计，被称为“部门”（sector），表10.4展示了分类系统分组的完整命名。运用6位数字可以让每个国家对其重要（却不一定对3个国家都重要）的经济活动进行分类，增加了系统的弹性。

表10.4 NACIS结构与命名

2 位数	部　门
3 位数	子部门
4 位数	行业组
5 位数	NACIS 行业
6 位数	国别

北美行业分类系统是如何改变编码结构的呢？表10.3显示，标准行业分类（SIC）仅有10个子类，在NAICS中则扩展到了24个。一些新部门代表了SIC子类的重要部分，使得SIC的子类有了进一步细分。另一些NAICS部门则合并了SIC较低层级的子类。最大的变化发生在经济的服务部门，随着时间的推移，服务部门在成长和变化。

表10.3　标准行业分类子类与北美行业分类部分的比较

SIC 子类	NAICS 编码	NAICS 部门
农、林、渔	11	农业、林业、渔业、狩猎
采掘	21	采掘
建筑	23	建筑
制造	31-33	制造
运输、交通、公共设施	22	公共设施
	48-49	运输与仓储
批发	42	批发
零售	44-45	零售
	72	餐饮服务
金融、保险、不动产	52	金融与保险
	53	不动产、租售
服务	51	信息
	54	专业、科学、技术服务
	56	行政、支持、垃圾管理以及矫治服务
	61	教育服务
	62	健康照料和社会支持

编码结构也有不同的变化方式。如此，到底选择怎样的逻辑呢？不管怎样，都得接受这样的目标，即NAICS应该反映当前的经济结构。不过，“经济结构”一词可以有不同的定义或解释，因此，行业分类系统可能基于一个或多个这样的概念。从经济理论（Economic Classification Policy Committee，1993）出发，至少有两种途径：需求分类系统和供给分类系统。需求或商品偏好的分类系统，其基础

是组织的产出。而供给或生产偏好的分类系统，则会把生产相同或相似产品的组织进行归类。生产偏好的编码结构意味着北美统计机构应该进行生产率、单位劳动成本、资本密集性的统计；构建投入-产出关系；估计就业产出关系，以及其他类似的同时运用投入和产出的统计。

标准行业分类是1930年代研发的，没有一个一致的概念框架。有些行业编码基于需求方，另一些则基于生产方（例如糖业，因生产过程不同，被纳入了3个不同的组；而乐器，尽管生产过程不同，却被归入了一个行业）。缺乏统一的分类概念也难以解释为什么采用这种方式分类而不是另一种方式。NACIS是基于生产偏好经济概念的。经济单位如生产产品或服务的过程就被归在一起了。

10.2.5　其他一般编码系统

在不同的领域，还有一些其他非常重要的分类框架。一些编码方案也运用到健康领域，如用于医生的医疗诊断，健康照料投资系统的管理，以及流行病统计。国际疾病分类（The International Classification of Disease，ICD）和国际疾病、诊疗分类（The International Classification of Disease，Clinical Modification，ICD-9-CM）就运用于死亡证书的编码和死因分类以及医疗记录的报告。精神疾病诊断和统计手册（The Diagnostic and Statistical Manual for Mental Disorders，DSM）则运用于精神流行病研究，搜集精神健康症状和针对具体症状应对的编码。一些健康调查也运用这些编码方案报告健康状态。

还有一些用于地理实体的编码和分类框架。每10年，美国管理和预算局（The U.S. Office of Management and Budget）要对大城市的区域进行一次确认，包括具体的人口和经济数据。此外，还有区域单位，被称为地面区块（tract），用于每10年一次的人口普查。每一次的人口普查，都假设这些有2 500～2 800人组成区块在人口特征、经济状态和生活条件上是同质的。每一个区块内，又划分为组块和块，两者都属于内部划分。在许多住户调查中，如果与人口普查的区块划分相匹配，则可以分析背景对家户或个人行为影响。这就要求对家户地址、组块、区块有编码。

最后，有些调查（如家户调查、农业调查）还运用了GPS（Global Positioning System）技术以标记样本要素，既用于快速定位，也用于未来的数据补充。GPS提供了地理位置信息，如此便可以把样本观察与其他空间（如遥感图、水资源，以及土地利用）数据联系起来。

如果通过合并和拆分把一个变量划分到不同的类，就可以把调查记录与其他数据源进行关联，对调查而言，是非常有帮助的。

10.2.6　编码中的质量指标

如果编码结构建构的质量不高和（或）不是统一执行的，则有可能在编码过程中产生误差。这里，我们讨论两类质量影响：（a）编码结构的弱点，（b）编码员方差。

编码结构的弱点。 如果为了分析，把两个并不等价的应答合并为一个编码类，可能导致一致性和系统性误差。例如，分析专家要测量大学学位和高中文凭或普通教育文凭（General Education

Development，GED）对薪酬影响的差别。GED指的是一个人通过了高中阶段多门课的考试后获得的文凭。不过，即使获得了GED文凭，也许在高中阶段的学习并不太好（譬如没有完全按照教学大纲执行），获得GED的人与正常高中毕业的人就不一定具有相同的知识，且在劳动力市场上也有不同的待遇。如果把高中毕业生和GED获得者的应答合并为一个编码，则在与大学毕业生比较薪酬时，其差别会大于把高中毕业生和GED获得者分开与大学毕业生进行比较的差别。

编码员方差。 在调查方法中，人们对编码质量的考察重点关注的是因编码员编码的变异性对调查估计值变异性带来的增加。回顾第9.3节对访员方差的讨论，两者之间有相同的逻辑，即编码员方差的形成与访员方差的形成非常相似。*编码员方差* （coder variance）是因访员运用编码结构的模式不同所产生的调查统计值整体方差的一部分。说的是，每一个编码员对编码结构的运用是有差异的，即运用编码分类的倾向性是不同的，把应答归入某个类以及运用剩余类编码（例如，“请具体说明”“其他”）的倾向并不相同。这类方差构成的量级通常用与测量访员方差相同的组内相关来测量，条件是给编码员随机分配编码样本。

表10.5展示了在英国做的编码员方差，ρ 值测量到的编码员效应平均为0.001。与第9.3节讨论的访员方差比较，编码员方差要小，不过，在讨论调查估计值方差时，还是要考虑到编码员方差。和访员方差一样，较小的编码员方差对调查统计值方差的影响为：

$$Deff = 1 + \rho_c (m - 1)(1 - r)$$

表10.5　职业编码中编码员方差统计

编码类	组内相关 ρ_c	信度估计	编码员设计效应
经理/行政主管岗位	0.005	0.881	1.02
专业人员岗位	−0.001 9	0.859	0.91
助理专业人员和技术人员岗位	0.001 8	0.836	1.11
行政和秘书岗位	0.003 4	0.935	1.09
手工操作及相关岗位	0.000 1	0.929	1.00
私人和看护服务岗位	0.002 5	0.950	1.04
营销和服饰服务岗位	−0.000 8	0.888	0.97
车间和机器操作岗位	0.000 0	0.904	1.00
其他岗位	0.003 1	0.943	1.06

数据来源：Source：Campanelli，Thomson，Moon，and Staples，1997；设计效应假设工作量为322个样本。

式中，ρ_c 指编码员组内相关；m 是每一个编码员编码的平均数量；r 是某个编码员的信度。

在实践中，编码员工作量往往比访员的工作量大。譬如，在表10.5中，每位编码员平均编码322个样本。这意味着平均设计效应为1+0.001（322-1）（1-0.903）=1.03，或在给定编码结构时，因编码员变异性，某个类别有3%的方差波动。

很少有研究探讨如何降低编码员操作的变异性，尽管培训可能是一项重要措施。Cantor和Esposito（1992）用定性方法研究了访员对行业和职业的编码，提出了改进编码质量的建议：

1）训练访员尽可能不过滤应答。

2）训练访员认识到获得职业名称的重要性。

3）如果列出多种活动，则允许访员追问。

4）为访员追问提供参考资料。

10.2.7　编码小结

简而言之，计算机辅助调查应用的一个意外后果是减少了对文本的编码。编码结构或多或少地反映了对调查访题应答进行解释的概念框架。为了用于分析，编码结构应该区分出在具体统计中具有区分度的应答。因此，有时候，针对同一道访题会采用不同的编码结构，产生不止一个分析变量。

运用给定的编码结构可能会带来变异性，由此，编码员有可能会增加调查估计值的不稳定性。训练有素的编码员产生的误差比访员产生的误差有效，不过，由于编码员工作量远大于访员工作量，有时候可能会增大标准误差。

10.3　录入

“数据抓取”（data capture）通常指把数值型数据输入到电子文档的过程。在实践中，数据抓取特别依赖数据搜集模式。在大多数计算机辅助调查中，要么是访员录入数据，要么是受访者录入数据。在按键电话输入和语音识别调查中，受访者直接将数据输入到电子文档。如果运用纸版问卷，则数据录入可能由录入员一份一份地输入，要么运用标记识别，要么运用光电扫描。

运用录入员手工录入是一项高成本的设计，正因为如此，人们渐渐地试图借助计算机辅助，以降低成本。手工录入一个操作特征是：

除了100%的录入，还要100%的校验。许多证据表明，如果是这样，则录入误差会很低。在1990年的美国人口普查中，误差率为0.6%（U.S. Bureau of the Census，1993），美国人口普查局的收入与项目参与调查，误差率为0.1%（Jabine，King，and Petroni，1990）。如此，尽管手工录入效果不错，不过其高额的成本还是激励调查研究者采用计算机辅助模式，进而消灭数据搜集后的手工数据录入。

10.4　清理

清理（editing）就是统计分析之前对搜集到的数据进行核查和修改。核查，是访员、督导、工作人员或专家在问卷访问完成之后进行的检查，也可以是计算机软件进行的检查。清理还可能涉及对单个测量或复合测量进行检验。清理的目的是对数据进行校验，看数据属性是不是与原始的测量设计一致。

清理渐渐地包括了访员或受访者为改善数据质量所进行的修改。有时候，清理还包括了编码和补值，即把缺损的数据补上。

完成清理要通过不同的核查。常用的核查包括：

1）阈值清理（range edits）（例如，年龄为1～120岁）。

2）比率清理（ratio edits）（例如，农场的牛奶生产量应该与奶牛数成比例）。

3）历史数据比较（例如，第二轮调查的家户人口数应该与第一轮的数有关）。

4）平衡清理 （balance edits）（例如，在家时间的比例与工作时间以及花在其他事项上面时间的比例加起来应该为100%）。

5）最高值、最低值以及其他不合理值核查。

6）一致性清理 （consistency edits）（例如，如果年龄小于12岁，则婚姻状态就应该是“未婚”）。

图10.2表明，计算机辅助访问软件已经把大多数清理纳入到了数据搜集中。这就要求受访者澄清并在理想状态下解决任何可能的问题。不过，Bethlehem（1998）注意到，要澄清误差有时候会很复杂（例如，一次涉及多个变量的不符）。此外，如果要解决清理失败问题，则访问时长就会增加，进而中断访问的风险就会提高。当然，也不是所有的清理都可以纳入计算机辅助访问应用程序中（例如，把受访者应答与外部数据进行比较）。最后，如果受访者坚持某种应答模式，且无法通过检查，则整个访问，根据研究者的逻辑，就不能反映受访者的真实情况。正因为如此，计算机辅助访问用户区分了必须遵循的“硬核查”和需要遵循的“软核查”（允许时候修改）。

不管样本是来自信息丰富的抽样框还是来自追踪调查，一项调查数据清理的工作量是实际搜集（逻辑一致的结构） 数据量的函数。正因为如此，从商业公司追踪搜集经济数据的调查通常会有大量的数据清理工作。譬如，1990年左右，针对95项联邦政府不同调查的研究表明，人们花费了大约20%的总预算进行数据清理，其中大多数都是有影响的调查。大多数调查在进行了某类自动或手工清理之后都有主题专家对数据进行评估。

如果组织得不好，则清理工作会伴随对数据无穷无尽的修改，由此数据质量也可能下降。例如，假设要核查年龄与受教育程度，却发现一位14岁的受访者有博士学位。一些看起来可能的事儿却真的不大可能。假设要核查年龄与职业，发现一位14岁的受访者不算劳动力。假设要核查年龄与户主的关系，发现一位14岁的受访者是户主的儿子。假设14岁的受访者的确是一位男性，则其受教育程度就可能有问题。如此，就要对受教育程度进行修改（不管是修改为缺损值还是依据14岁男性与父母居住的状态补值）。在看到下一个变量之前，这一改动似乎很好。假设有一道访题“您发表过任何作品吗？”，应答是“是的，博士论文”。如此，则原有的应答记录是对的。如果是这样，难道年龄记录是错的？有没有可能是稀有事件呢，即一个人14岁获得了博士学位？如果没有系统的规则指导这类决策，就极有可能无法迭代出一个解决方案，即使是清理者自己，几周之后也不可能重复原先的决策。

Fellegi和Holt（1976）发明了一个整合清理与补值的系统，其中包括一整套步骤，按照这些清理规则就可以重复。这是一项重要的贡献。这项技术始于三个判断：

1）清理应该用最少的变量值修改让所有数据通过检验。

2）清理应尽可能保持边缘和联合频数不变。

3）补值规则应该来自于清理规则的迭代。

任何清理，不管原始状态如何，可以被分解为一系列的表述形态，但“不可以有一个具体的编码组合”。例如，一个双变量记录：

年龄（计算到上一个生日的整岁）

婚姻状态

1）未婚。

2）已婚。

3）离婚。

4）丧偶。

5）分居。

一项明确的清理规则是，如果受访者小于12岁，则应记为“未婚”。这就意味着，应答“小于12岁”与其他婚姻状态的组合都不成立。如此，则可以表述为：

（AGE＜12且MARSTAT=已婚）=不成立or

（AGE＜12）∩（MARSTAT=已婚）=不成立

（AGE＜12）∩（MARSTAT=离婚）=不成立

（AGE＜12）∩（MARSTAT=丧偶）=不成立

（AGE＜12）∩（MARSTAT=分居）=不成立

这个清理规则意味着可以派生出其他变量的清理规则。人们可以从显性清理派生出隐性清理。**显性清理**（explicit edits）指研究者

在清理时，每一个调查记录都必须符合的规则。隐性清理 （implied edits）是类似的规则，逻辑上暗示着要满足显性清理的规则。例如，假设一个数据记录有年龄，则针对受访者是否登记投票，是否投过票，有下列两条清理规则：

（AGE＜18）∩（REGISVOTE=是 ）=不成立

（REGISVOTE=否） ∩（VOTED=是 ）=不成立

隐含着另一条规则：

（AGE＜18）∩（VOTED=是 ）=不成立

Fellegi-Holt方法通过识别数据记录中需要清理的大量数据的字段而管用。通过某种修改进而让数据通过审核。由于过程是事先设定的，故一旦修改，在显性清理和隐性清理都会通过审核。原理是，通过清理不成立频次的逐步核查而实现。用于清理的软件系统已经开发出来了，主要用在政府统计中，运用的也是Fellegi-Holt方法。

清理小结。对清理系统的应有特征，人们有一些共识，包括与概念测量联系在一起的显性规则；结果的可重复性；向基于规则的、计算机辅助的、更节省的清理活动转化；尽量少地改动问卷数据；把清理与补值结合起来；以及清理结束的标准化即所有记录都通过审核。未来的清理将与过去不同。随着计算机辅助在调查环节的前移，清理系统会变成与调查其他步骤整合的活动。数据搜集完成后的清理工作

会减少。还有可能的是，软件系统开发也会越来越多地与研究主题的专家合作。

10.5　加权

数据搜集之后，还有一个步骤，即使是大量采用计算机辅助调查也要做的，就是为进行统计分析而加权。

运用复杂抽样方法获得的调查样本具有不同的备选概率、应答变异性以及关键变量与已知外部数据分布比较时的偏离。为此，对复杂抽样的调查要进行加权，以弥补这些问题。

调查抽样中的加权有不同的来源和背景。在这里讨论的目的，主要是例举复杂调查中常用的调整方法。除了这里讨论的方法以外，还有其他的加权方法。这里的主要目的是例举，不是全面介绍（参见Kalton，1981；Bethlehem，2002）。

这里将列举复杂调查中用到的4种加权方法：

1）用第一阶段比例调整 （first-stage ratio adjustment）的加权。

2）为不同入选概率加权。

3）为因样本无应答加权。

4）用于降低抽样方差（覆盖不足以及问卷无应答）的事后分层加权。

10.5.1 用第一阶段比例调整加权

在分层多阶段抽样中，譬如NCVS和NSDUH，初级抽样单位（PSUs）通常会依与规模成比例的概率抽取。抽样框会说明比例，通常与目标总体规模或代表目标总体规模指标的成比例。

在等概率抽样设计中，每一层抽选的样本单位数应该与该层目标总体的规模成比例。假设NCVS的县层有该县0.5%的家户人口，据此可分配初级抽样单位和次级抽样单位的多阶段区域概率样本。如果抽中某县，则初级抽样单位的层总体规模为：

层总体规模估计 =PSU总体/抽选 PSU的概率

为降低抽样方差，第一阶段比例调整就是对选中PSU的所有样本加权，

W_{i1} =第一阶段比例调整权重

=从抽样框获得的层总体/（PSU总体/抽选PSU的概率）

这里做的就是稳定初级抽样单位的估计值，对所有样本加权就是实现抽样设计的一致性（Cochran，1977）。

对PSU的每个受访者，会生成一个与第一阶段比例调整权重等价的新变量，即W_{i1}，下标“1”表示权重因素中的第一个加权因素，大写字母用于提示。这个权重是基于框总体的权重，而不是基于样本数

据的权重。最终，在第10.5.4节呈现完每一个常用加权因素后，就要将其综合为一个最后的权重，即来自于单个因素的汇集。

10.5.2　为不同抽选概率加权

假设在12岁及以上群体中要抽取125 000个个人样本，假设抽选的范围为美国2.85亿人口，其中有1.995亿或70%的人口年龄在12岁及以上，则总体抽样份额为f =125 000/199 500 000=1/1 596。

拉丁裔人口的增长以及拉丁裔和非拉丁裔人口犯罪受害者的差异便提出了这样一个问题，在125 000个样本中，拉丁裔样本数足够吗？假设12岁及以上人口中，拉丁裔占1/8，或总人口有约2 500万人，如按等概率原则抽样，则样本中应该有15 625个拉丁裔以及109 275个非拉丁裔。也就是说，依据比例分配原则，在样本中也应该有1/8的拉丁裔样本。或许抽样结果中拉丁裔的比例就是1/8，如此则无需进行加权。

假设情况不是这样，拉丁裔的样本数占到总样本量的一半，即62 500。相应地，非拉丁裔的样本数就会下降，因为总样本量依然是125 000。与成比例的样本配置比较，如果特别希望估计拉丁裔受害者的状况，这样的样本分配就具有吸引力。

为了实现这样的样本分布，针对拉丁裔的抽样比例就得大幅提高，从1/1 596提高至1/399。与此同时，非拉丁裔的比例就得下降至1/2 793。如此，拉丁裔和非拉丁裔的样本数均为62 500。

在后面的计算中，只要把拉丁裔组和非拉丁裔组分开统计，这样的抽样是没有问题的。如果要合并计算，问题就来了。如果要统计全国的状况，而不是区分族群，譬如计算女性受害者状况，就需要合并计算。

如果不区分族群，针对总人口进行估计，则需要解决拉丁裔过度代表的问题。对个体值加权后进行计算就是实现这类调整的一种方式。如果有个体权重，则加权平均值为：

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_{i2} y_i}{\sum_{i=1}^{n} w_{i2}}$$

正如在第4.5节分层抽样讨论中提到的那样，可以运用备选概率的倒数作为每个样本的抽样权重 （selection weight）。我们用w_{i2}代替w_i作为权重中的第二个权重，第一阶段比例调整是第一个权重。样本中的每一个拉丁裔的权重为399，非拉丁裔的权重则为2 793。

在权重均值（$\bar{y}_w$）的计算中，权重（w_{i2}）既出现在分子中，也出现在分母中，则从代数的观点看，无论权重大小，估计值应该不变。这就是说，对均值而言，重要的不是权重的绝对值，而是相对值。如此，权重399和2 793应该可以转化为更易记忆和核查的数值。譬如2 793/399=7，如此，针对所有拉丁裔，加权1；针对非拉丁裔，则加权7。

对两组而言，这个权重很大。在合并计算时，拉丁裔数值的分布将会降低至非拉丁裔的1/7。如此，在总体中就更正了样本对估计值的贡献。

在调查数据集中，每一个拉丁裔的权重为1（或399），同时，非拉丁裔的权重则为7（或2 793）。在合并计算时，这些权重就会产生适宜的补偿。在分组计算时，也可以使用权重，且在比较时不会产生偏差，因为同一组的每一个样本加权的是同一个值。

10.5.3 为因样本无应答加权

类似于NCVS这样的调查，会出现无应答。无应答的比例，在每一个组都可能不同。还是以NCVS为例，年轻一些的组（如12～44岁组），应答率为80%；年长一些的组（45岁及以上），应答率为88%。则125 000样本中的应答样本为105 000，其分布见表10.6。

表10.6 假设给拉丁裔和非拉丁裔分配等量样本的无应答调整

	总体规模	样本规模	应答数	应答率	无应答调整权重 w_{i3}	无应答调整后的权重
拉丁裔	24 937 500	62 500	52 500	0.84		
12~44 岁		31 250	25 000	0.80	1.25	1.25
45 岁及以上		31 250	27 500	0.88	1.14	1.14
非拉丁裔	174 562 500	62 500	52 500	0.84		
12~44 岁		31 250	25 000	0.80	1.25	8.75
45 岁及以上		31 250	27 500	0.88	1.14	7.95
合　计	199 500 000	125 000	105 000			

这个分布假设样本（应答的和无应答的）年龄（至少年龄组）为已知。对无应答的调整，这是一项重要约束。在实践中，对无应答进

行调整的前提是，可以获得每一个样本要调整的变量构成。也就是说，如果只知道应答者的相关变量，而不知道无应答者的相关变量，就无法进行调整。

在初始样本中，年龄在12～44岁的样本数与45岁及以上的样本数是相当的。不过，在应答者中，45岁及以上的样本数要大。如此，就过度代表了年长群。

为调整过度代表，我们可以假设调查中的无应答加权（nonresponse weight）就相当于曾经讨论过的不等概率抽样的加权。假设一个子群（这里就是年龄组）应答者是受访者的随机样本，那么，应答率就相当于抽样比率。如此，可以假设无应答是随机的，也是对无应答进行加权的基础。无应答率的倒数就可以作为权重，让受访者分布等于初始样本分布。

这些调整权重，加上基础权重，用于调整不等概率抽样。对拉丁裔而言，基础权重为1.0，非拉丁裔的基础权重为7.0，正如表10.6所示，无应答调整权重则来自基础权重（即$w_{i1} \times w_{i2}$）和无应答调整。

如果把无应答调整权重运用到105 000个受访者，则权重和就等于两个年龄组之和。换句话说，通过加权分布，让每个年龄组的人数相等。

10.5.4　后分层加权

在许多调查中，最后一个加权程序就是后分层。假设通过无应答调整，分男女对加权后的每个性别求和，且和相等。同时，通过外部数据可知，在总体中，女性比男性的人数要多，即52%对48%。**加权后分层**（poststratification weight）运用权重以确保样本与目标总体来源的某种分布一致。无论是对统计效率还是对展示而言，让每个性别的权重与外部来源的分布一致，都是有帮助的。要做到这一点，就可以对无应答调整权重做进一步的调整。

这里指的是，把男性的权重降低，把女性的权重提高。如果我们把每一位男性受访者的权重降低，即（0.48/0.50）=0.96，把每一位女性受访者的权重提高，即（0.52/0.50）=1.04，则外部总体的性别分布与加权后的样本分布就一致了。在表10.7的倒数第二列，后分层加权权重用w_{i4}表示，因为它是第4个权重因素。运用大写字母，意味着其基础是目标总体，而非样本信息。

表10.7　加权后的样本分布及后分层（假设的NCVS样本，区分性别、年龄、种族）

	受访者	加权后无应答和	加权样本分布	总体分布	后分层权重 w_{i4}	最后权重 $w_{i1}\times w_{i2}\times w_{i3}\times w_{i4}$
男性	52 500	250 000	0.50	0.48		
12~44 岁	25 000	125 000				
拉丁裔	12 500	15 625			0.96	1.20
非拉丁裔	12 500	109 375			0.96	8.40
45 岁及以上	27 500	125 000				
拉丁裔	13 750	15 625			0.96	1.09
非拉丁裔	13 750	109 375			0.96	7.64
女性	52 500	250 000	0.50	0.52		
12~44 岁	25 000	125 000				
拉丁裔	12 500	15 625			1.04	1.3
非拉丁裔	12 500	109 375			1.04	9.1
45 岁及以上	27 500	125 000				
拉丁裔	13 750	15 625			1.04	1.18
非拉丁裔	13 750	109 375			1.04	8.27
合　计	105 000		105 000			

10.5.5　权重的汇集

最后，要获得权重是把第一阶段比例调整（w_{i1}），不等概率抽样调整（w_{i2}），无应答调整（w_{i3}），以及后分层（w_{i4}）合在一起。最终的产出是把4个权重（$w_{i1} \times w_{i2} \times w_{i3} \times w_{i4}$）分别给予8类的每一类：男性12～44岁拉丁裔组，男性35岁及以上拉丁裔组，等等。表10.7列出了每一类的最后权重。例如，12 500个拉丁裔12～44岁组的最后权重是1.20。

这个最后权重综合了4种不同的权重，在分析计算时，应该加载到每个个案数据上。一些数据集会把每个权重作为一个单独的变量，结果也是一样的。

在数据分析中，是在任何情况下都要运用权重，还是只在某些情况下使用权重，是一个问题。之所以有这样的问题，是因为会有许多估计值，加权或不加权差别不大。如果的确如此，为什么要加权呢？特别是在明确知道加权会导致方差增大、结果精确性降低的情况下。

有一些分析证明了加权是不必要的。例如，在比较如表10.4所示的小样本数据时，如果组内每个样本的权重是一样的，则无需加权了。

如果是分层抽样，则不同层的样本权重原本就不一样。尽管对某个具体变量而言，是否加权并无区别，却不能保证对其他估计值而言也没有差别。在需要对子类、样本子集进行分别估计时，尤其不知道“无差别”的条件是什么。

Ekholm和Laaksonen论对样本无应答进行倾向值加权

Ekholm和Laaksonen（1991）用应答倾向值的期望值对应答进行加权，弥补了样本无应答。

研究设计：芬兰家户预算调查运用了对总体进行分层随机抽样的个人样本。调查要估计总消费、每个家户不同物品和服务消费的平均值。总应答率为70%。作者运用了基于家户构成的4因素（城市、区域、资产收入以及从128项跨类合并）Logistic回归来预测应答者的似然值。当把叠加单元格与拟合不佳的模型关联以后，计算了125个单元格的每个单元格的倾向值，并与运用35个后分层权重计算的结果进行了比较。

研究发现：估计应答概率为0.40～0.90，最低的是居住在大城市区域的老年单身家户，无资产收入；最好的是居住在中等区域的家有两位年轻成员的家户，有一些资产收入。运用倾向值对所有个体加权，比后分层加权的估计结果要好。因为，倾向值小的家户通常人口少、消费也少，用倾向值加权，则平均每户的消费均值也低。

研究局限：对用倾向值加权的估计值没有外部效度。没有不同做法的备选倾向值模型。

研究影响：研究显示估计每一位受访者应答概率的多变量Logistic模型可以用于样本无应答的补偿。

分析者有时候需要用权重，有时候则不需要。如此，就给解释或解读留下了困难。

调查统计学家在对调查数据的分析中通常会用权重。在对总体及其子类的描述性统计中，还会纳入设计特征，如不等概率抽样和无应答，分析性统计与描述性统计一样。

10.6　为选项缺失值补值

上一节对无应答调整的讨论强调了无应答的一种形式：样本无应答。对访题无应答不一定适用。

正如在第2.2.8节和第6.6节讨论的那样，访题无应答意味着在访问或问卷中，在其他变量获得数据的情况下，一些变量却没有获得数据。例如，在NCVS中，一位受访者在受访中提供了刑事犯罪受害者信息，却不一定愿意提供家庭收入的信息。

图10.4（也出现在第2章中）汇集了调查中的两类无应答。这个图是一个矩形的调查数据集，行为样本数据，列为变量数据。数据集后半部分为样本无应答，只在左边有数据，即抽样框数据。如此，数据集是按应答状态排序的。即使是样本无应答，从抽样过程也会获得一些样本信息。在类似于NCVS的调查中，至少有家户位置等样本信息。

图10.4　调查数据文件中的样本和选题无应答

在受访者记录（图的上部）中，在数据集中也有用一个数字代表的访题无应答。变量不同，访题无应答的比例也不相同。在调查中，无应答比例最高的通常是收入和资产类的访题。

如图10.4所示，数据集有缺损数据是数据分析遇到的最突出问题。在分析中，如何处理缺损数据才是最好的方法，我们并不知道。常见的有两种方法：“忽略”问题，或对数据进行调整。

一种策略就是把有缺损值的样本删除，或用完整样本数据进行分析。如果在分析中把有缺损值的所有样本数据删除，就是忽略访题缺失值。在NCVS调查中，如果用多个预测变量（如年龄、性别、家庭收入）与被抢劫变量做回归分析，那么，在分析时，这些变量任何有缺

损值的样本都会被剔除。也就是说，如果受访者没有应答是否被抢过，没有提供年龄或性别信息，或拒绝提供家庭收入信息，则整个样本都会在分析中被剔除。

从推论的角度看，用调查数据推论总体，并运用完整数据进行分析，就是一种调整形态。剔除的样本并非真的被忽略了。运用完整数据进行分析的背后，实际隐含着对缺损值的补值或替换。如此，完整数据分析用平均值或完整数据的结果“补偿”或指定了缺损值的值。换句话说，缺损值并没有被忽略，在推论中，数据分析者用应答数据替代了无应答数据。

另一种让数据完整以便于分析的方法就是补值。补值（imputation）就是用估计应答填补到缺损值、错误值、野码值位置上。当然，任何调查机构采用补值法补值，都会对补值变量进行标记，以便于分析者确定在分析中是否接受补值。

补值方法有优点，即知道补值方法，分析者每次使用的也是同一个值，还有，未缺损的访题数据也用上了。同样也有缺点。对许多分析者而言，所补之值，无论补得多么好，都是“假”值，当然，分析者无法忽略的是，运用完整样本数据分析时，即使忽略有缺损值的样本，依然存在着隐含补值。此外，几乎所有统计分析软件都会把补值当真值处理。如此，标准误估计值就会被低估，进而导致置信区间过窄或检验值过大。

为弥补缺损值，补值有众多的程序。这里只讨论其中的极少数。

也许最简单的方法就是填补。假设某个变量有缺损值，则可以用这一变量应答值的平均值譬如说$\bar{y}_r$ 填补缺损值。例如在NCVS中，家庭收入的缺损值就用所有报告了家庭收入的平均家庭收入（或中位数，或众数，如果家庭收入有分类）进行填补。用平均值补值大致相当于用完成的数据进行分析。

平均值补值 （mean value imputation）有一些缺点。如果在补值数据集中有许多缺损值，则变量值的分布就会出现“尖峰”。如此，补值后的数据分布就被扭曲了。

对扭曲有多种处理方法。第一，为补值加上一个随机要素，以反映估计值的已知变异性。如此，补值便成为了$\bar{y}_r+s_i^2$ ，这里s_i^2 也许是从正态分布中选取的均值为0、方差等于该变量非缺损值方差的值。

也可以用子类的平均值（或许更准确）补值。例如，在NCVS中，可以分子类如种族或族群进行家庭收入补值。用已知数据可以计算出非洲裔美国人和非非洲裔美国人子类的家庭收入，如此，非洲裔美国人家庭收入的缺损值就用非洲裔美国人的家庭收入平均值进行填补。同样也可以为这个值加上随机值，以避免扭曲子类家庭收入的分布。

这种子类补值方法，加上一个随机产生的残差，可以被认为是对缺损值的回归预估。如果模型中有种族变量，则家庭收入是二分变量x_1 的回归值。

$$y_{i(r)}=\beta_0+\beta_1 x_{1i(r)}+\varepsilon_{i(r)}$$

在估计中，仅仅考虑了访题应答（用下标r 表示），则对每一个缺损值有

$$y'_{i(r)} = \beta'_0 + \beta'_1 x_{1i(r)} + \varepsilon'_{i(r)}$$

式中，β'_0 和β'_1 为估计系数，估计残差则为加载预测值上的随机残差。

回归补值程序也可以用于多于一个预测变量、二分变量与其他预测变量混合的情形。这就是说，对一个缺损值可以依据已有应答值进行回归补值：

$$y_{i(r)} = \beta_0 + \beta_1 x_{1i(r)} + \beta_2 x_{2i(r)} + \cdots + \beta_p x_{pi(r)} + \varepsilon_{i(r)}$$

对p 的预测变量x_j 而言，j =1，2，…，p 。

在调查中，回归补值有时候也用于对几个变量的补值，也必须符合补值变量值的分布。例如，如果家庭收入是分类变量，就得用多个分类Logit模型获取预测概率，并回过来转化为有随机分布或无随机分布的某个类。

回归补值 （regression imputation）是用所有已有值作为预测变量来预测缺损值（因变量）。回归补值或许还要求进行顺序补值，如此让补上的值也可以派上用场。

回归方法有许多有价值的特征，不过对大规模补值而言，人们通常使用几十年前开发的一种方法：顺序热值方法。*热值补值* （hot-deck procedure）也可以被看作回归补值的一种形式，只是其值是从数据集的另一个样本借的。补值示例参见表10.8。

表10.8 顺序热值补值示例（家庭收入、补值、补值标记变量）

受访者序号	性别	受教育程度	家庭收入	热值	补值数据	补值标记
1	M	9	23	51	23	0
4	M	11		23	23	1
2	M	12		23	23	1
3	M	12	43	23	43	0
7	M	12	35	43	35	0
8	M	12	42	35	42	0
5	M	16	75	42	75	0
6	M	16	88	75	88	0
16	F	10		88	88	1
15	F	12	28	88	28	0
17	F	12	31	28	31	0
18	F	12	35	31	35	0
19	F	12	30	35	30	0
22	F	12		30	30	1
13	F	14	67	30	67	0
14	F	15	56	67	56	0
21	F	15	72	56	72	0
20	F	18	66	72	66	0

假设有一个小数据集需要对家庭收入补值，对数据集的18个样本而言，已知性别和受教育程度，其中有4个样本的家庭收入值缺损，也就是说有家庭收入数据的有14个样本。顺序热值补值从对数据集按照性别和受教育程度排序开始。表10.8的数据先按照性别排序，然后再按受教育程度排序。如此，让相同性别、相似受教育程度的样本数据集中在一起。

补值，从观察排好序的数据开始，看家庭收入的第一个数据，如果缺损，就用全部样本或子样本的家庭收入均值填补。在示例中，有应答的男性家庭收入均值为51 000美元，如果第一个样本的家庭收入值缺损，就用这个值填补。类似于这样的初始补值，通常又被称为冷值（cold-deck value）补值。

如果家庭收入的第一个样本值没有缺损，则受访者报告的值就是热值，也就接着往下看。如果下一个值缺损，就用上一个热值填补。在表中，第二值缺损，就用第一个已有的值填补。如果第二值没有缺损，则它就是下一个样本的热值。

以此类推，再看下一个值。如果缺损，就用热值填补。如果没有缺损，则它就是下一个样本的热值。如此，用最近的一个热值对缺损值进行填补。由于相邻的两个样本在与收入相关的变量上相似，补值也正是把这些排过序的变量当作了回归补值的预测变量。还有，如果排序要补值的样本值是可识别的，譬如说，有报告值，就可以认为，补值是用具有相同应答值的样本平均值加上了残差。此时，残差来自于排序后有数据的另一个样本的值。

在实践中，还有一些其他的补值方法，不过，内容已经超出了本书的范围。正如之前指出的那样，补值后的数据会低估估计值的方差。不过，这个低估，也可以通过特定的方差估计程序来弥补，要么通过补值过程本身，要么通过多次补值。多次补值（multiple imputation）会产生多个补过值的数据集，每一个数据集来自于每个补值模型的不同实现方法（参见Rubin，1987；Little，and Rubin，2002）。多个数据集的估计值的变异性就可以用来估计总变异性，包括抽样方差和补值方差。

10.7　对复杂样本的抽样方差估计

分层多阶段抽样、加权、补值是调查数据的特征，且要求运用非标准程序正确地估计估计值的方差。正如第4.5节所示，要分层首先必须对层内方差进行估计，并把不同层的方差进行综合。

调查分析者总是在寻求把这些设计特征与具体的方差估计程序和软件应用结合起来。如此，形成了三类针对调查数据某个特征的、可以运用软件的方差估计程序。这里，做一简要说明。

运用分层多阶段抽样并通过加权来弥补不等概率抽样和无应答的调查数据集，有时被称为*复杂调查数据*（complex survey data）。估计这类数据的抽样方差在多种可用软件中有三个过程可用：泰勒序列近似，平衡重复复制，或者杰克剪刀重复复制。

泰勒序列近似（taylor series approximation）是一种常用程序，用于不能简单地为样本值进行方差估计的情形。例如，比值比的方差就很难获得，因为比值分母中的倾向或样本规模就是样本估计值本身。

泰勒序列近似在处理这类难题的时候，是把比值转化为近似值，这个近似值不包含比值，却是样本值和的函数。泰勒序列近似用在许多统计中，也用于有权重的分层多阶段抽样设计。例如，一个简单比值均值的方差，

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

运用泰勒序列近似（为了简单起见，假设样本来自于简单随机抽样），则

$$\frac{1}{(\sum w_i)^2}[Var(\sum w_i y_i)+\bar{y}_w^2 Var(\sum w_i)-2\bar{y}_w Cov(\sum w_i y_i,\sum w_i)]$$

这个式子看起来复杂，其实只是简单随机样本方差估计不同计算的合并。在复杂抽样设计中，对均值、比例抽样方差的估计，泰勒序列近似也许是最常用的方法。因为，大多数统计软件都包含了这一方法。

平衡重复复制（balanced repeated replication）和杰克剪刀重复复制（jackknife repeated replication）采用的是完全不同的方法。有别于为抽样方差估计寻找分析途径，这两种方法有赖于重复子群抽样。人们可以想象，重复复制就相当于从同一个总体同时抽取不是一个而是许多个样本。对每一个样本而言，都有一些统计估计值，譬如均值$\bar{y}_\gamma$ 就来自于对每一个样本 γ 的计算。

对总体均值的估计则可以被看作不同样本估计值c 的均值，

$$\bar{y}=\left(\frac{1}{c}\right)\sum_{\gamma=1}^{c}\bar{y}_\gamma$$

这个估计值方法是每一个单独样本估计值围绕均值的变异，或者，

$$v(\bar{y})=\left(\frac{1}{c(c-1)}\right)\sum_{\gamma=1}^{c}(\bar{y}_\gamma-\bar{y})^2$$

这个方法的长处是可以把估计统计值抽样方差的方法应用于几乎所有统计值、均值、比例、回归系数以及中位数。

平衡重复复制和杰克剪刀重复复制过程则基于稍有不同的方法。对两者而言，都用全样本中可识别的子样本测量变异性。这些子样本，通常被称为“重复”（replicates），抽取自原始样本，如此，每一个重复都携带着全样本的基本特征（除了样本要素编码数）。例如，如果样本来自整群和群的分层，则子样本抽样过程就是从每一层抽取群样本。如此，就可以得到许多子样本重复，子样本估计值的变异性就是估计总样本估计值抽样方差的基础。

用第一个子样本就可以计算重复估计值$\bar{y}_\gamma$ 。然后从样本中抽取第二个子样本，再进行估计。如此抽取子样本或复制过程，会多次重复。然后用与前述非常相似的过程计算估计值的方差。不过，在方差估计中，不是运用估计值的均值，而是用总样本的均值。

平衡方法与杰克剪刀方法的区别在于子样本的抽取方法。典型的平衡重复复制过程被称为“半样本复制”，即抽取一半的样本或一半的群样本或一半的其他什么样本。杰克剪刀重复复制过程则又被称为“剔除”过程，通过从数据集中剔除一个样本、一个群，进而获得每一个子样本。

这些方法因具体目的在不同的调查估计软件中有一些变种。把这些软件的名字列出来并不意味着支持某个软件、不支持另一些软件。它们都需要层、群或初级抽样单位，加上用于估计的权重。如果是分层随机抽样，就可以用样本层级的识别变量，而不是群或PSU。如果无需加权，就可以让变量的权重等于1，进而用于分析，且不会对结果造成影响。如果不分层，就可以让所有样本的层变量等于1，然后进行计

算。为了估计方差，所有软件都要求每一层至少有两个群，或要素，或PSU。

CENVAR是美国人口普查局开发和发布的一个统计包，是一个微机软件。CENVAR可以运用泰勒序列近似方法估计样本和子群的均值、比例方差。软件为菜单式，可以处理分层要素、分层多阶段样本设计。这个软件包是免费的，可以从人口普查局的国家项目中心获得（详情请参见http://www.census.gov）。

VPLX也是美国人口普查局开发的一个软件包，它比CENVAR的应用范围要宽。如果重复有权重，则可以进行后分层处理。VPLX运用重复复制过程，包括平衡和杰克剪刀方法。类似的还有CPLX，用于列联表分析。VPLX和CPLX都是免费软件，可以从人口普查局网站上获得。

EpiInfo则是另一个软件，由美国疾控中心开发和发布。EpiInfo可用于复杂抽样调查数据的方差估计。它运用了泰勒序列近似方法进行总样本和子样本的均值与比例方差估计。EpiInfo用于视窗（Windows）系统，可以在疾控中心的网站下载（http://www.cdc.gov/epiinfo）。

接下来介绍的软件，都是收费软件，不可以免费下载。

SAS 9，一个运用广泛的统计软件，有多个程序（PROCs）可用于复杂抽样设计的方差估计。其中一个程序可以处理基本均值和比例；另一个程序则处理线性回归，方差估计则采用了泰勒序列近似。关于SAS的进一步信息可以参见http://www.sas.com。

STATA是一个可编程统计软件（参见http://www.stata.com），包括了众多的统计程序，其中的一些程序也可以用于复杂抽样调查数

据。较新的程序叫作“svy”或与调查有关的各种命令。统计程序则包括了描述性统计如均值、比例、线性、logstic、多个定类变量logit、probit、Cox比例风险以及泊松回归。

SUDAAN是研究三角机构（Research Triangle Institute，详情请参见http://www.rti.org）开发的软件，用于复杂抽样设计方差估计。研究者可以选择泰勒序列或重复复制程序，此外还有许多分析方法，从描述性统计到线性、logstic、多个定类变量logit、probit、Cox比例风险的比值估计。和SAS相似，运用关键词调用程序，语法也与SAS相似。此外，SUDAAN的程序也与SAS兼容，可以直接在SAS上运行。

最后，WesWar是西部统计（Wesstat，Inc，详情请参见http://www.westat.com）开发的一个统计软件，用于复杂抽样数据的重复复制方差估计。软件可以用于多种估计，从和、均值、描述统计到线性回归，也可以用于后分层效应估计。它的老版本为免费版本，最新版本则是收费版本。

10.8　调查数据的文档与元数据

人们很少只是为了一个分析者建构数据集，也很少有调查数据集只是用于一次分析，更少有调查数据在完成数据搜集之后就用于分析，且此后就束之高阁。与之相反，调查数据常常会为许多人用于分析，且分析很多次，用很多年。的确，大型调查数据档案，类似于ICPSR（Interuniversity Consortium for Political and Social Research，http://www.icpsr.org），或Roper

Center（http://www.ropercenter.uconn.edu）为分析者搜集了几千个调查数据集，有些数据甚至是20世纪早期的。

正是在这样的条件下，调查设计者必须考虑到数据要为许多人所用，不仅是项目人员，而是全世界的人员。这就要求与数据集相关的文档能够让人们在分析数据之前理解数据。这类关于数据的信息（数据的数据），就是元数据。*元数据*（meta data）描述的是任何潜在数据用户为有效利用数据所需要的信息。如果数据用户不设限，则对需要什么样的元数据也没有定义。在调查背景下，Colledge和Boyko（2000）列出了常见的元数据：

1）定义性的——描述目标总体、抽样框、访题措辞以及编码术语。

2）过程性的——描述访员培训、获取抽样框、选择受访者、搜集数据的方法。

3）操作性的——对过程的评估，譬如访题数据缺损率，清理失败率，访问平均时长，以及每位访员完成的访问数。

4）系统性的——涉及数据集格式、文档位置、使用协议、变量标签与说明、变量类型等。

对元数据类型的这种说法，与早期非电子数据文档、用纸版编码簿（codebook）、用备忘录方式描述抽样和调查设计等标准方法的元数据大相径庭。不过，基础编码簿依然是常规文档的一部分。

例如，图10.5就是2001年NCVS数据编码簿的一部分。请注意，在数据文件中，每个变量都有不同的输入。图中显示了两个字段：一个

是VAR V2081，是一道破坏财物的过滤访题；另一个是VAR V2082，则是受访者报告损坏了什么。每个字段都有一个字段头用于定义信息（变量名和访题措辞），以及系统信息（数据记录位置，如COL140WID1，即字段的开始位和字段宽度；缺损数据编码值，数据集的层级，即家户层级还是个人层级数据；还有非缺损数据的编码值与定义）。对V2082，还有一些用大写字母给出的过程元数据“标记在给定时间内损毁的所有财物”，这是给NCVS访员的说明，告诉访员如何记录受访者的应答。最后，对V2802而言，还有一些信息涉及搜集数据之后的数据清理、再编码。（受访者对多个访题的应答输入，详情列在VARS V2083-V2092。）这些信息提醒用户这些数据字段来自于调查的后处理过程，综合了某个受访者应答的综合测量。编码簿对分析者而言是一个关键工具，它是数据搜集过程与单个数据之间的桥梁。

VAR V2081 VANDALISM AGAINST HOUSEHOLD NUMERIC
COL 140 WID 1 MISSING 9 HOUSEHOLD DATA

Source code 557

Q.46a Now I'd like to ask about vandalism that may have been committed during the last 6 months against your household. Vandalism is the deliberate, intentional damage to or destruction of property. Examples are breaking windows, slashing tires, and painting graffiti on walls. Since ______ __, 19__, has anyone intentionally damaged or destroyed property owned by you or someone else in your household? (Exclude any damage done in conjunction with incidents already mentioned.)

1. Yes
2. No
3. Refused
8. Residue
9. Out of universe

VAR V2082 LI VANDALISM OBJECT DAMAGED NUMERIC
COL 141 WID 1 MISSING 9 HOUSEHOLD DATA

Source code 558

Q.46b What kind of property was damaged or destroyed in this/these act(s) of vandalism? Anything else?
MARK (X) ALL PROPERTY THAT WAS DAMAGED OR DESTROYED BY VANDALISM DURING REFERENCE PERIOD.

Lead-in code

(Summary of single response entries for multiple response question. Detailed responses are given in VARS V2083-V2092)

1. At least one good entry in one or more of the answer category codes 1–9
8. No good entry (out of range) in any of the answer category codes 1–9
9. Out of universe

Note: For a "Yes–NA" entry, the lead-in code is equal to 1, the category codes are equal to 0 and the residue code is equal to 8.

图10.5　全国刑事犯罪受害者调查编码簿示例

随着从纸版编码簿转向电子调查文档，元数据在过去的几年里得到了极大的丰富。随着超链接能力的增强，片段化的元数据可以连在一起，同时提供不同层级的累积信息。这个领域的研究前景与发展可

以让用户直接运用互联网通过多种方式对数据提出需求。如果用户是初学者，他可以要求了解某个具体领域（如社会经济地位）的测量方法，并列出一组变量。这组变量也许可以链接到每个变量的元数据：访题措辞，既有研究对访题的利用，已存储的应答方差，之前文献对数据的分析和利用，甚至传统元数据用到的编码、缺损值、数据记录位置。这类元数据结构可以预测大量用户类型的需求，并在信息之间建立链接，进而尽快地从调查数据中获得信息。

这一趋势为调查方法专家在元数据领域提出了前所未有的问题，负担与机会并存。调查方法专家有了大量展示调查质量测量的机会，用于产生简单应答估计方差的重访调查数据可以很容易地提供给用户。与行为编码、认知访谈以及其他问卷开发相关的评估的多次前测访题，可以在使用应用中直接链接到最后的访题上。在用户分析时，所有涉及测量质量的信息都可以关联起来。这个新世界的负担就是调查设计必须把元数据设计纳入其中。在讨论基本概念、目标总体、测量以及抽样议题时，设计者必须考虑什么信息可以让结果数据集适应多样化用户的需求。

10.9　小结

在调查数据搜集完成之后，并非马上可以用于分析。不过，到底要经过多少步骤才可以让数据用于分析，则取决于数据搜集模式。计算机辅助减少了一些原本需要在数据搜集完成之后的步骤。

清理数据是一个清洁数据、消除明显错误的过程。随着数据搜集中计算机辅助的增加，数据清理也被整合进数据搜集之中。每一个不同的应用或许都需要相应的编码结构。在编码步骤也会带进统计误

差，在编码过程中的编码员变异性既会带来统计量的系统误差，也会让精确性降低。与访员方差的测量相比，对编码员误差的定量估计相对容易。

在统计分析中对数据进行不同的加权是为了抵消抽样的不等概率，调整抽样框中的缺失，调整样本无应答，以及通过运用目标总体的已知信息来改善估计精度。基于样本数据的加权通常会增加调查数据的标准误，也希望因此降低由无覆盖和无应答带来的统计偏差。运用了框总体属性的加权则一般会降低标准误。

补值是在统计估计中用某个值替代缺损值的过程，通常也是访题层级的过程，而不是样本层级的过程。补值是为了降低访题无应答带来的偏差。运用有关缺损观察值的丰富辅助数据，就能达到这一目的。补值重复运用了受访者的数据，尽管基于补值数据的估计值标准误要高，数据集则相对完整。

运用调查数据集估计抽样方差可以让研究者反映由设计带来的可测量的变量误差源。估计抽样方差有多种相互替代的方法。标准统计软件在逐步加入的常规统计步骤有分层、整群以及调查中运用的不同权重方法。

调查文档和元数据是一项调查至关重要的产品，为了让这类数据可用，在抽样设计时就要认真考虑。随着万维网和超链接关联相关文档能力的增强，调查文档的特征也在快速变化。

关键词

编码（coding）

实地编码（field coding）

清理（editing）

比率清理（ratio edits）

一致性清理（consistency edits）

隐形清理（implied edits）

抽样权重（selection weight）

分层后加权（poststratification weight）

平均值补值（mean value imputation）

热值补值（hot-deck procedure）

复杂调查数据（complex survey data）

平衡重复复制（balanced repeated replication）

元数据（meta data）

编码结构（code structure）

编码员方差（coder variance）

阈值清理（range edits）

平衡清理（balance edits）

显性清理（explicit edits）

第一阶段比例调整（first-stage ratio adjustment）

无应答加权（nonresponse weight）

补值（imputation）

回归补值（regression imputation）

多次补值（multiple imputation）

泰勒序列近似（taylor series approximation）

杰克剪刀重复复制（jackknife repeated replication）

编码簿（codebook）

进一步阅读资料

Biemer, P., and Lyberg, L. (2003), *Introduction to Survey Quality*, New York: Wiley.

Lyberg L, Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.) (1997), *Survey Measurement and Process Quality*, New York: Wiley.

Särndal, C.E., and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, New York: Wiley.

作业

---

1. 针对口头应答，与数据搜集完成之后再进行编码比较，请说明实地编码的一个优势和一个劣势。

2. 在一项针对使用汽车旅行的旅游调查中，您正在核查每年行驶里程的录入数据，在2 000个调查样本中，您遇到了三种情况：

   （a）有一个样本，受访者报告数据为17 500英里，录入数据为1 750英里。

   （b）另一个样本，受访者报告数据为17 599英里，录入数据为17 588英里。

   （c）第三个样本，受访者报告数据为17 599英里，录入数据为15 799英里。

   在单独分析中，得到的全样本里程均值为15 004英里。在识别这类错误用到了如下质量保障技术：

   （a）核查关键数据的分布，并识别野码值。

   （b）抽取10%的样本进行核查，估计并修正误差。

   （c）进行100%重录入，并进行校验、修正误差。

   请说明每一项技术在消灭录入错误中的成本与收益。

3. 运用您已经掌握的知识，说明建构权重以用于调查数据估计的一般步骤。

（a）界定什么是“抽样权重”均值。

（b）说明基于权重的每一个调整步骤。

针对每个调整步骤，说明权重带来的影响。

4. 说明运用应答均值作为访题缺损值补值的两个问题。

5. 考虑下面的样本，n =20，来自于概率抽样，并有基础权重用于调整不等抽样概率。

（a）计算未加权的慢病数量均值和针对不等概率抽样加权后的慢病数量均值。比较两个均值，比较并说明其异同。

（b）根据近期的人口普查数据，可以得到下列W_{i4}的数值：男性40岁以下，0.22；40岁或以上，0.24；女性，40岁以下，0.22；40岁或以上，0.32。在基础权重之上，计算每一组的后分层调整因素，并获得一个加权分布，让其与普查数据一致。

（c）在后分层过程中，哪几类调查误差需要调整？

基础权重	性别	年龄	慢性病数量
4.4	M	24	0
4.2	M	37	1
2.4	M	66	2
3.0	M	57	3
2.0	M	23	2
2.4	M	26	0
2.6	F	28	1
3.0	F	32	1
3.0	F	39	0
1.1	F	40	1
1.3	F	41	0
1.2	F	47	2
1.4	F	43	1
1.5	F	48	1
1.1	F	53	3
1.4	F	38	3
1.8	F	63	4
1.1	F	68	1
1.1	F	73	2
1.0	F	78	3

6. 您已经完成了对下列变量的测量，依据应答产出，假设某州驾照的核发起始年龄为16岁（这里“是否有驾照”指的就是这种条件下）：构建这5个变量的数据清理列表，并核查变量之间的逻辑。

性别	是否有驾照	年龄	就业状况	开车作为临时职业
女性	有	<16岁	有工作	是
男性	无	16岁或以上	学生	否
			没工作	不适用(学生或没工作)

7. 从概念上说明平衡重复复制和杰克剪刀方法在抽样方差估计上与泰勒序列近似有什么不同?

8. 简要说明热值法，与应答均值补值比较，说明其优势和劣势。

9. 什么是多重补值？多重补值与单次补值比较，其优势如何？

10. 访问NSDUH和BRFSS网站。

（a）说明每一项调查的后分层过程。

（b）每一项调查中，哪一些估计最容易受后分层因素的影响?

11 研究伦理的原则与实践

11.1 导言

与前面的章节不同，这一章的内容与调查结果的统计误差属性没有直接关系，而是讨论大多数社会对研究，特别是涉及对人的研究进行考量的背景。讨论指导调查方法专家进行选择的规范与规则。最后，对规则涉及的调查设计的调查方法研究进行回顾。

与调查研究有关的伦理实践，可以大致区分为两类：第一类涉及研究实施的标准，即调查专业人员提供的操作意见。第二类是研究者对其他人员应尽的责任，特别是对受访者、客户以及公众的责任。

这一章将简要涉及研究实施的标准，以及研究者对客户和公众的责任。不过，重点还是放在研究者与受访者的关系上，即两者之间的伦理议题上。只有目标总体的友好才能让我们完成工作，因此，保护受访者的利益的确事关我们自己的利益。

11.2 实施研究的标准

和一般的研究者一样，调查研究者也遵循科学活动的一般原则。正如在第1章提到过的，调查方法是一项科学探索的领域，也是更大的调查研究的一部分。在国际上，有许多调查专业组织，其中WAPOR，即世界公众舆论研究学会；ESOMAR，世界市场研究组织就是例子。在美

国，最能代表调查研究者的，是美国公众舆论研究学会（AAPOR），它纳入了学术、商业、政府等部门的研究者。AAPOR的会员有责任遵循学会的伦理准则（http://www.aapor.org），它是少有的几个提供对违规处罚机制的工具，这里我们就用其作为例子。总体上看，它要求会员在从事研究时要遵循准则，并采用必要的步骤保证结果的效度和信度。不过，1996年5月针对调查研究领域标准实施的最佳实践与AAPOR谴责的实践之间曾经有过一场争论。会员们还是坚持了前者，遏制了后者。

最佳实践指的是运用科学方法抽样以保证每个总体成员有可测量的入选机会，追随入选样本以获得足够的应答率，仔细开发和测试问卷以了解访题措辞和顺序对应答带来的影响，以及对访员以足够的训练和督导。这本书按章节顺序为这类最佳实践提供了详细的指导，这里将不再重复具体的内容，而只讨论与调查研究伦理相关的议题。除非能获得有效的结论，否则要求受访者参与进而给受访者带来负担就有悖伦理。

在APPOR界定的“不可接受”（尽管类似于最佳实践列表不是AAPOR伦理准则的正式部分）包括在研究的外表下进行的筹资、销售、拉票活动，把自选或志愿应答者的民意测验结果（例如读者偶尔反馈的问卷，拨打800或900电话投票，或者在网络问卷上偶尔填的问卷）作为有效调查的产出进行展示，以及为进行投票为受访者提供虚假或误导信息的偏向性民意测验。例如，一项偏向性民意测验或许会询问受访者，如果知道了候选人X因虐待儿童而被起诉、对宠物不友好等，是否会投反对票。询问这类访题的动机不是为了发现针对这类议题的公众舆论，而是为了在选民中散布不信任、不支持的种子。在没有获

得研究参与者许可的前提下披露其可识别信息，也是不可接受的调查研究实践。

和其他科学家一样，调查研究者也秉持科学活动的一般标准。在美国，联邦执行部门为大多数人类主题（其中大多数涉及生物、医学）提供资助的是健康与人类服务部（Department of Health and Human Services）。在这个部门中，研究综合办公室（Office of Research Integrity，ORI）负责监管科学研究的不当活动，包括剽窃（plagiarism）、作弊 （falsification），或在申请、执行、评审，以及报告研究结果中的造假 （fabrication）。这间办公室对这些数据都有清晰的界定，参见表11.1。

表11.1　研究活动不当的关键术语

术语	定义
造假	捏造数据或结果，并进行报告。
作弊	操纵研究素材或过程，改变或删除结果，致使研究并非准确代表研究本身。
剽窃	偷窃或滥用和大量复制不属于自己的他人知识产权。包括未经许可特定的交流，不包括署名权或名誉权滥用。

涉及科学和非科学领域研究行为不当的可靠信息非常少，很难判定的是，偶尔由报纸报道的事件到底是冰山的一角，还是极少的例外（Hansen and Hansen，1995）。就像研究一般越轨行为一样，对研究活动不当的研究非常困难，也涉及无应答误差和测量误差。不过，这个领域的一些近期研究表明，研究活动不当或许比科学家们愿意相信

的更加普遍（Swazey，Anderson，and Lewis，1993；Martinson，Anderson，and De Vries，2005；Titus，Wells，and Rhoades，2008）。

此外，尽管对访员作弊的关注不多，却很持续。访员作弊（interviewer falsification）指有意偏离访员指南或说明且不做说明的行为，它可能导致数据污染。“有意”意味着访员知道自己的行为偏离了指南和说明的要求。

作弊包括：

1）捏造全部或部分访问数据，即记录的数据并非来自研究设计指定的受访者，且将其作为受访者数据报告。

2）有意误报处理码造成程序数据作弊（例如把拒访记录为样本不合格；编造联系次数）。

3）为跳过接续访题而有意对应答进行错误编码。

4）为减少完成访问的工作量而访问非样本对象，或向调查管理部门有意歪曲数据搜集过程。

在所有大型调查中，如果访员是临时人员且没有作为研究进程的一部分，就有可能出现作弊。另一方面，作弊的比例的确不高。对作弊进行研究的文献也很少，表11.2列出了美国人口普查局的研究结果（Schreiner，Pennie，and Newbrough，1988）。在这项研究中，作弊包括在样本家户中选择非样本个体作为受访对象，或用非授权的方式选择受访对象。表中的结果显示，与当前人口调查（Current

Population Survey）和全国刑事犯罪受害者调查（NCVS）这样的持续调查比较，在类似于住房空置调查（Housing Vacancy Survey）的一次性数据搜集中，这个比例较高。

表11.2 美国人口普查局主持的三项调查中访员作弊的比例

调查	作弊百分比
当前人口调查	0.4%
全国刑事犯罪受害者调查	0.4%
纽约市住房空置调查	6.5%

Schreiner，Pennie和Newbrough（1988）也发现，与有经验的访员比较，经验较少的访员更倾向于作弊，而有经验的访员作弊的模式则更加复杂（例如在追踪调查的第一轮调查中作弊）。Schaefer，Schraepler，Mueller和Wagner（2005）对近期的研究进行了综述，Smith等（2004）则对调查中如何预防作弊进行了探讨。

11.3 对待客户的标准

依据AAPOR伦理规则，针对客户的伦理行为，第一，鉴于调查研究技术本身的局限性或研究者或客户资源的局限性，要求研究者仅接受可完成的研究委托。第二，除非有客户直接许可或有更高层级规则的要求，否则，研究者必须对研究数据严格守密。最重要的是，研究者只要知道研究发现中有数据扭曲，就要公开披露，包括对将要提交结果的群体，不管需要通过怎样的努力来矫正这些扭曲。在实践中，这

些条款让研究者有一种义务，如果他们知道有扭曲的结果发布了，就需要对管理机构、管理者、媒体以及其他相关群体提供矫正说明。可以理解的是，如果客户依据调查或民意测验发布了扭曲的结果，依据这些条款，就会与研究者制造利益冲突。

一个著名的例子是，Roper Organization的负责人Bud Roper了解到，在询问纳粹是否真的实施过对犹太人的灭绝时，受访者的某个应答比例极高。这是Roper Organization接受美国犹太人委员会委托进行的一项调查，由于问题措辞为："如果说纳粹对犹太人的灭绝从未发生过，您认为可能还是不可能"，在获得的应答中，有22%的受访者认为纳粹对犹太人的灭绝并未真实发生过，出现这个比例是完全可能的。为了检验这个结果，Roper用自己的经费对修订过的访题进行了重访，并公开修订了调查结果。为此，Roper说，"最终，我们用自己的钱做了重访，全样本重访，因为这道题变了。您猜，我们获得了什么？认为大屠杀未曾发生过的比例只有1%"（参见Krazit，2001）。这样的行为对一个人的声誉、经费都意味着高昂成本，的确也超出了准则的要求。根据准则要求，Roper只需要一纸说明，承认可能有错。

11.4　对待公众的标准

AAPOR的伦理准则试图通过准许对公众披露民意调查或调查结果最低要求的信息来履行职业责任。

表11.3列出了AAPOR 8项的要求，包括说明调查的资助方和调查地点，还包括与调查结果的不同误差来源有关的关键设计特征。

表11.3　最低限度披露内容（AAPOR）

1. 谁是调查资助方，谁是调查执行方。
2. 访题的具体措辞，包括可能影响应答的任何指导语、说明语。
3. 研究总体的界定与抽样框说明。
4. 抽样过程。
5. 样本量，如果可能，完访率，以及合格样本信息，还有过滤样本的过程。
6. 准确的发现，包括：如何合适，样本误差估计值，权重或估计过程。
7. 如果有，依据部分而不是全样本获得的结果。
8. 数据搜集的方法、地点、日期。

数据来源：http://www.aapor.org。

1）覆盖性误差——研究者必须说明目标总体和抽样框。

2）抽样误差——研究者必须说明抽样设计和样本规模。

3）无应答误差——研究者必须说明完访率（应答率的一个部分）。

4）测量误差——研究者必须说明数据搜集方法、问卷和指导语的措辞。

如此，在最低限度上，这些信息传递了某种观点，一项调查的估计误差，或反过来说，调查结果的可信程度与这些误差有关。伦理准则要求，任何研究结果报告，或在发布时，都必须有这些内容。换句话说，伦理准则尽管没有要求研究者们追随最佳实践，正如本书讨论的，却要求披露实践的过程。因此，伦理准则试图在“思想市场”上不断驱逐最糟实践。

作为披露实践的一个例子，在SOC的网站（http://www.sca.isr.umich.edu），读者可以找到对调查的一般说

明、抽样设计的细致说明，还有问卷，以及调查关键统计量的计算说明。

11.5　对待受访者的标准

11.5.1　对受访者应承担的法律义务

在美国，以人（包括调查受访者）作为研究对象的法律基础是1974年的研究法案（P.L. 93-348，July 12，1974）。加拿大和澳大利亚针对以人为对象进行研究的法律与美国类似。在加拿大，所有涉及以人为对象的研究都必须经过研究伦理委员会（Research Ethics Boards，http://www.nserc.ca/institution/mou_sch2_e.htm）的审查；在澳大利亚，类似的机构被称为人类研究伦理委员会（Human Research Ethics Committees，http://www.health.gov.au/）。在欧洲，除生物医学领域之外，似乎没有任何规则来约束以人为对象的研究活动。不过，在欧盟却有规则保护个人数据的安全性（http://www.coe.int）。

在美国，1974年的《国家研究法案》催生了“保护作为研究对象的人类法”（联邦法，45 CFR 46 May 30，1974，46.3©），法案的最近一次修订是1991年。依据这些法律，学院、大学以及其他接受联邦资助的机构，都有建立伦理审查委员会以保障研究志愿者包括调查受访者的权利。伦理审查委员会（Institutional Review Boards，IRBs）是由研究者、地方社区代表组成的委员会，负责审查以人为对象的研究申请，确认被研究对象的权利是否获得了充分的保障。

根据“联邦保护作为研究对象的人类法”，如果记录的信息可以直接或间接地识别出研究对象，研究对象的应答如果被披露会有损受访者的声誉、职业活动以及经济信誉，或者让被研究对象可能面对民事或刑事指控，依据45 CFR 46.101，则IRBs审查就会通不过。不过，近些年，因志愿者在一些著名大学临床实验中死亡而引发的公众愤怒，政府关闭了一些大学参与的研究，IRBs也收紧了对研究项目的审查，越来越少地运用联邦法案提供的促进研究或通过评审促进研究的机会（Citro，Ilgen，and Marrett，2003）。一个全国性的趋势是，由人类研究保护项目委托委员会（Association for the Accreditation of Human Research Protection Programs，AAHRP）提供的委托或许可以缓和这样的格局。

Tuskegee梅毒研究

在美国公共卫生服务局的资助下，Tuskegee雇用了携带梅毒的南部黑人男性做追踪研究以观察疾病的阶段。这项研究从1932年开始，那时还没有针对梅毒的有效治疗方法，一直由政府方面的科学家在继续，即使有了盘尼西林，即有了新的、有效的治疗方法，也没有告诉研究对象。事实上，被研究者从一开始就被欺骗，让他们相信他们会得到治疗，事实上，却并没有给他们治疗。

这样研究雇用了大约600人，全部都是非裔美国人。其中，399人有梅毒，却被告知他们的血液不好，在当地，即意味着贫血。虽然给他们提供治疗，使用的却是安慰剂。不少人失明，甚至罹患精神病。

这项研究于1972年终止。直到1993年，美国政府都没有为这项研究正式承认错误或向受害者道歉。Tuskegee研究主要是政府在非裔美国人身上的医疗失信。

目前，只有受联邦政府资助的机构进行的调查项目才要遵循“保护作为研究对象的人类法”。因此，大多数商业调查是不受监管的（包括，如知情同意）。不过，这样的格局或许会改变。2002年，参众两院的立法或许会把在美国进行的所有研究纳入，无论经费从哪里来，都要接受人类研究保护办公室（Office of Human Research Protection）的监管。不过，这项法案的执行，在商业性调查中还有很长的路要走。

11.5.2 对受访者应承担的伦理义务

伦理有别于法律，用于保护受访者权利的原则来自于Belmont报告（National Commission for the Protection of Human Subjects of Biomedical Research，1979）。1979年，美国专门针对把人作为生物医学和行为研究对象进行保护的人类法委员会发布了这份报告。这个委员会是根据1974年的国家研究法案成立的。

在美国，保护作为研究对象的人类这一运动可以溯源自在生物医学研究中对研究对象权利的侵害，特别是纳粹时期的德国科学家、美国Tuskegee梅毒研究中的科学家（Faden and Beauchamp，1986；Katz，1972；Tuskegee Syphilis Study Ad Hoc Advisory Panel，1973）以及其他拿患者进行医学实验的研究。

尽管早期对人类权利的侵害发生在生物医学研究领域，一些社会科学研究（如Laud Humphreys在公共厕所对同性性行为的观察，Humphreys，1970）和社会心理学研究也涉嫌欺骗（如Zimbardo对模拟监狱的研究，Haney，Banks，and Zimbardo，1973，参见http://www.prisonexp.org/）也引起了公众对潜在伤害的关注。Milgram（1963）的一项非常重要的对权威服从的研究涉嫌为希特勒领导下的德国人暴行辩护，他招募“科学家”的“合作者”对看不见的志愿者进行电击。科学家（即Milgram的助手）根据被电击者（也是Milgram的助手）的反馈，不管被电击者如何哭泣，都不断提高电击强度。事实上，并没有任何电击，实验的真正被试是（相信自己）实施（真正）电击的人，实验观察的是他们表现出的、在权威指导下给人带来痛苦的不同意愿。当然，在实验结束后，当Milgram及其助手在对实验做了简要说明后，他们也经历了不同的苦闷。

多重顾虑导致了1974年《国家研究法案》以及同年的联邦法案细则与适用情境的出台。1991年，17家联邦机构重组构造了45 CFR 46的子模块A，也被称为一般准则。

Belmont报告对所有涉及以人作为对象的研究而言有3个领先的原则，即对对象：善行、公正、尊重。

善行（beneficence）原则要求研究者为研究对象带来尽可能小的伤害和尽可能大的收益，并权衡在某种风险下获取收益是合适的或在什么情况下因风险而让收益尽失。联邦法案的细则更多地关注到风险与伤害正反映了善行原则。

公正（justice）原则指让研究带来的负担与收益之间有一个公平的平衡。例如，在19世纪和20世纪早期，贫困的病人承受了医学研

究的大部分负担，而由医疗改进带来的收益却带给了那些付费病人。在心理学实验中过分依赖心理学学生，除了研究结果的一般化可疑之外，也可以被看作违背了公平原则（转引自Rosenthal and Rosnow，1975）。

第三个原则对人的尊重 （respect for persons）提出了一项伦理要求，即知情同意。知情同意 （informed consent）可以被定义为“个体或监护人……在没有过分诱导，或任何强制、强迫、欺骗、谎言，或任何形式的限制或约束下，知道且同意”（U.S. Department of Health，Education，and Welfare，1974，p. 18917）。

正如Smith（1979）在他的文章“社会科学研究的一些伦理和政治视点”中指出的那样，对研究对象的伦理和法律责任在于对人的善行与尊重（来自于不同的哲学体系），相互之间并非一定互适与融洽。史密斯就其中之一论述到：

> 自由主义、唯意志论、“人道主义”框架只是抓到了“知情同意”的一个标签……如果人们能自己做主，那的确不错，问题是他们不得不做出类似自我牺牲的选择，如果结果对他们自己有害。如果他们被强制、被欺诈或被操纵，即使对他们自己有利，也是要不得的。其他框架，如对参与者有利还是有害，显然来自于“功利主义”传统……知情同意，也非来自于自由意志假设，听起来似乎更有客观性，实际可能是时髦的成本收益分析的一种表达。

Smith接着讨论了由此给社会科学研究带来的问题，即后面我们将进一步讨论的尊重与善行伦理原则。

11.5.3　知情同意：对人的尊重

很多人认为知情同意是为了保护作为研究对象的人（包括受访者）免受伤害。可争论的结果是，如果没有伤害的风险，就没有必要知情同意。不过，如前所述，知情同意的真正目的是让受访者和研究对象对自己的信息具有控制权，即使没有伤害问题。例如，即使可能招致某些对象拒访，告知潜在受访者其参与访谈的自愿性对调查实践的伦理而言依然极其重要。同样的，即使最终不会给受访者带来伤害，在获得受访者许可的前提下进行记录也是标准要求。目前，不少州都要求，只有获得受访者同意才可以对访问进行记录。

表11.4列出了共识规则下对知情同意的关键要求，包括对研究目的的说明，参与调查可能带来的收益与伤害，数据保密申明，以及参与的自愿性。

表11.4　知情同意的关键要求

1.	一份申明，说明研究的目的，预计研究对象参与的时长、过程以及任何实验性过程的识别。
2.	说明任何可能的风险或不适。
3.	说明研究对象或其他人可以期待的任何收益。
4.	说明备选过程或处理。
5.	说明如何保密以防止通过记录识别对象的身份。
6.	如果研究有可能带来伤害，则说明如果发生伤害会给予什么，以及怎样的补偿或处理。
7.	说明在有进一步涉及研究、研究对象权利、研究相关的伤害问题时可以跟谁联系。
8.	说明研究对象是自愿参与的，且在任何时候都可以退出参与且不受惩罚，利益不受损失。

在具体情况下，伦理审查委员会可能会免除部分甚至全部要求，甚至完全免除知情同意。

在生物医学领域，知情同意的沟通媒介和文档都是书面的，且要求有受访者的签字。在以下两种情况下，伦理审查委员会可以免除对受访者签字的要求：（1）在研究中，与研究对象唯一相连的记录就是知情文档，且主要风险来自于违背保密原则；（2）研究呈现的风险不大于最小伤害，且在研究背景之外不含书面知情同意过程。

许多甚至大多数调查都具备第二项特征，不过，有些显然不具备。例如，调查中询问到非法或污蔑行为（如NSDUH）就可能让参与者受伤害，如果有意或无意地披露信息。在受访者参与之前，他们有权知道这样的风险，还有，因潜在利益冲突，研究者不能单方决定向受访者披露多少以及以何种形式披露信息。在这种情况下，伦理委员会就应该要求用书面文档向受访者描述风险，且说明如何避免风险以保护研究对象，以及在受到伤害之后如何救助。伦理委员会也应该要求在受访者参与之前实际拿到了文档。不过，这类文档是否以受访者书面签署形式，则是另一个问题。

尽管一些主要专业学会的伦理规则（例如美国统计学会，美国社会学会）提到了知情同意要求，不过，AAPOR却没有，只是说“我们应该尽量避免可能对受访者带来伤害、羞辱，或严重误导的实践或方法”。对这样的缺失，并不难理解。与生物医学研究不同，调查质量依赖于应答率。如此，对受访者合作的需求尤其重要，在调查说明中又很难转达知情同意的要素，如果要求书面签署知情同意，就会降低应答率（Singer，1978，2003）。

此外，Presser（1994）认为，如果在调查中运用知情同意，就会使得与受访者的沟通更加艰难，也不可能预测任何准确性。例如，从伦理的观点看，需要清楚说明调查资助方和调查目的，同样，也会导致调查应答的偏差（Groves，Presser，and Dipko，2004）。如果说

明了，不仅一些人会不参与某些组织或某种目的的调查，即使参与了调查，也可能会因反对或迎合其所理解的调查目的而扭曲地应答。从另一方面来看，到底跟受访者说多少才合适。例如，是不是告诉受访者在访问结束前将询问其收入状况，或让受访者知道，如果他们不想回答，就可以不回答？

大都市项目

1986年，一组瑞典研究者开始了一个大都市项目，确认1953年出生的学龄儿童全体为样本。搜集的数据包括出生证明，与孩子父母的访谈，基于调查的循环问卷，以及一些人口登记信息。这项研究的一个目的是理解社会环境与健康状态的关系。

这样研究持续搜集数据，直到1980年代中期。不过，许多调查对象并不知道他们参与了研究，因为出访调查在他们还是孩子的时候就已经完成了。后续测量是整个调查的一部分，其陈述的调查目的与大都市项目无关。不仅如此，他们也不知道自己的人口登记信息在定期地更新到研究数据中。1986年2月10日，斯德哥尔摩的一家报纸《每日新闻》（Dagens Nyheter）刊登了一篇通讯，并强调了没有知情同意。

由此，展开了一项公开辩论，许多人，包括国会议员，对无视研究对象的权利，对没有说明如何搜集、整理、利用研究对象的数据，表示了愤慨。之后，政府要求对数据做匿名化处理，并约束未来的数据关联。（在公共辩论期间，与大都市项目完全无关的调查项目，应答率急剧下降。）

11.5.4　善行：保护受访者不受伤害

对调查受访者而言，或许最大的伤害风险来自于其应答的披露，即有意或无意地违背了保密原则。越是敏感信息，例如，HIV状态或是否有小的违法行为，违背保密原则所带来的伤害就会越大。因此，当要求受访者提供此类信息时，他们强烈要求保密就毫不意外了（Singer，Von Thurn，and Miller，1995）。

对调查信息保密的威胁是多方面的。最常见的，也是最简单的，就是大意：没有抹除问卷或电子文档上的可识别信息，或者对包含可识别信息的文件没有加密。尽管还没有证据表明这类忽视给受访者带来了伤害，对搜集数据的机构而言，提醒自己的员工重视这类问题、提供行为指导并保证这类提醒的可见性，是重要的。

不太常见、潜在的却更严重的威胁来自于合法地对识别信息的要求，无论是传讯还是引用信息自由法案（FOIA）。调查搜集到的非法行为数据如吸毒，潜在地具有被司法机构传讯的特征。为避免被传讯进而伤害受访者，研究者在类似NSDUH的项目如研究精神健康、酗酒、吸毒以及其他敏感议题的项目中，无论是否是政府资助的，都有可以援引健康与人类服务部的保密证（certificate of confidentiality）。这类证明在整个研究期间都有效，可以保护研究者在大多数情况下不用给联邦、州或地方部门（42 CFR 2a.7，“保密证生效”）被迫提供受访者的姓名或其他识别特征。

对统计数据的保密性（confidentiality）的额外保护还来自于2002年签署的法案。1971年，联邦统计局总统委员会建议“运用‘保密’条款总是意味着披露信息或以其他任何方式让公众可识别受访者

进而构成对受访者的伤害，在任何情况下，都是禁止的”，因此，这类“数据具有司法豁免性”。委员会还建议“司法部门应该授权为统计目的搜集数据且保证数据的保密性”。自此以后，才有不断地努力来加强统计信息保密性的法律保护，以及允许机构之间为了统计目的的有限数据分享。2002年12月，“保密信息保护和统计效率法案”终于让这类立法进入了法律系统。不过，2001年生效的国土安全法案让保密承诺最终得到解决，正如“谢尔比修正案”（Shelby Amendment，PL 105-277，1998年10月21日签署）的生效一样，原本用于研究的数据也被用于政策和规则制定。

如果是为了诉讼（而不是统计运用）而调查，从一方或另一方要求数据，不适用于运用“保密证”或新生效的法律。Presser（1994）回顾了一些案例。在这种情况下，研究者除非把数据藏起来，否则别无他法，即，要么违背保密承诺，要么进监狱。Presser（1994）讨论的许多行动，专业调查机构都可以拿来保护类似情形下的保密性对象，例如，正如一些调查机构在卷入诉讼中已经做到的那样，为了不至于反溯出识别信息而委托在调查中销毁识别路径。如果在调查产业中这种做法变成了标准操作，个体研究者就能更好地为自己的做法辩护了。

数据保密性的最后一个威胁来自于所谓的“统计披露”。统计披露（statistical disclosure）是指拿调查数据与外部数据进行匹配进而识别出个体或个体属性。到目前为止，在这一章讨论的都是受访者显性的识别特征，譬如姓名、地址、社会保险号。不过，人们逐步地了解到，即使没有显性的识别特征，运用高速计算机、外部数据包含的姓名与地址以及个体特征的多种信息，加上复杂匹配软件来匹配调查数据，进而识别受访者。追踪调查数据尤其容易匹配。

技术让识别受访者变得容易，或让反识别变得困难，也越来越多地与调查数据关联。这些技术，如DNA样本、生物测量特征以及地理空间特征，使得数据的匿名性变得更加复杂，也加强了研究者试图既让数据可用又加强保密的两难。现在，除了严格特殊许可、连带、特殊安排或创造综合的、插值的数据文件以外，似乎没有可接受的方法来保护包含可识别变量的保密性。稍后，在11.8.2节，我们再继续讨论。

由于上述对数据保密性的多重威胁，在研究者中越来越多地达成了共识，即很难有对受访者而言绝对安全的保证（National Research Council，1993）。譬如，美国统计学会的“统计实践伦理指南”提醒统计学家们要“预计到对研究主题数据的二次和间接使用”。至少，“指南”敦促统计学家们提供之后的、外部的、独立的、重复分析。“指南”也提醒研究者们不要用法律程序来隐含对保密性的保护，除非得到明确的授权。对敏感信息的搜集更要努力获得保密证，如此才能对受访者获得最强的保证。不过，即使有保密证也很难保证不会有统计披露。

11.5.5　说服时的努力

有一个议题值得在这里做简短的讨论，那就是希望那些在家时间不多、很少有整块时间以及很少答应合作的人群有高一些的应答率。这个两难让调查机构试图尽量避免拒访和变卦，运用尽可能大的激励促成应答。参与者接到后续的联系，试图说服他们应答却没有酬劳，以及给那些一开始就拒绝应答的酬劳，这二者在参与者看来都是令人感到不快的（Groves，Singer，Coming，and Bowers，1999）。通过

付钱让拒访者应答显然是不公平的，不过，这类事情似乎并不会降低未来应答意愿的表达以及新调查的参与（Singer，Groves，and Coming，1999）。

不过，从伦理的角度来看，这类实践引出了一些问题，请看下面的例子。

> 例子一：一个承包人为调查老年人获得了数百万美元的支持。他承诺了80%的应答率。在实地调查快结束的时候，应答率还是不到80%，此时，研究者让访员给可能应答的人100美元以促成他们接受访问。
>
> 例子二：在一项有关基因测试知识和态度的调查中，访问对象是大夫和基因咨询顾问。为增加应答率，研究者们决定给大夫们每半小时支付25美元。鉴于在试调查中基因咨询顾问的应答率略高，故不给他们支付酬劳。

两个案例提出了一个平等问题，不合作的人却要因获得其合作而付费，合作的人却要因其合作而无酬。（经济学家对这类问题有不同的观点，认为对应答的人，即使没有激励，调查的效用也较大，因此无需给予额外的激励。）

不过，在方法上的确是一个问题。例如有证据表明，运用特别方式进入调查的，大多数都是通过付酬转变的拒访者，正如第一个例子中展示的，与合作的、不是通过付酬转变的受访者比较，通过付酬的转变者通常更加富有，有了他们，样本数据更准确（Juster and Suzman，1995）。其他的研究也表明，金钱激励可以补偿受访者与研

究议题的无关性，进而在降低无应答偏差上非常重要（Groves，Couper，Presser，Singer，Tourangeau，Piani Acosta，and Nelson，2006）。

另一个问题是，激励的大小或类型在什么时候以及什么情境下可以让潜在受访者无法拒绝应答。例如，让囚犯减刑以换取其合作，或给吸毒者现金激励，或让因支付问题而得不到治疗的病患获得治疗以让其参与药物实验。这里有一个问题，用特殊的激励让受访者合作是否有悖自愿原则，即实际上是激励产生的强制性效果（参见Singer and Bossarte，2006；Singer and Couper，2008）。如果采用这些激励，伦理委员会审查极有可能通不过。

11.6　出现的伦理问题

在过去的若干年，运用互联网络进行调查的势头在增长，这类调查提出一类全新的伦理问题或增加了旧问题的新维度。例如，对隐私的保护和保密需要技术上的支持（例如“安全”网站或应答加密），超出了老的数据搜集模式。另一些伦理议题在老的调查模式中根本就不会出现，例如受访者多次应答，进而导致偏差。对研究者而言，这就是一个伦理问题，因为根据AAPOR准则，要尽量避免结论的误导性。还有一些其他的伦理问题，譬如没有做知情同意，在网站上做调查，既带来了新的挑战（如，如何确认没有未成年人参与，如何对记录键盘行为的记录做知情同意），也带来了新的便利（如互动知情过程，在这个过程中，受访者在进入下一个模块之前，必须确认已经理解）。对理解的讨论及其伦理问题，请参见相关文献（American

Psychological Association，2003；Singer and Couper，未刊稿）。

11.7　研究调查中的伦理问题

正如已经注意到的，调查研究者面对的两个最重要的伦理议题或许是在知情同意的前提下让受访者参与，以及对应答的保密，因为违反保密性最有可能让受访者受害。在这一节，我们讨论在实践中这些议题如何影响了调查参与。

11.7.1　研究知情同意协议

第11.5.3节讨论了调查中知情同意的关键部分。正如不少研究已经指出的那样，在调查实践中执行这些原则并不简单。在实践中，与方法研究有关的有三个领域：

1）研究受访者对知情同意内容的反馈。

2）研究方法调查中知情同意，特别是涉及的一些欺骗类型。

3）研究书面与口头知情同意，特别是请受访者签署的效果。

研究受访者对知情同意内容的反馈　Singer（1978）提供了一份折半样本，采用了两种内容描述方式（参见文本框 ）。对其中的一半运用简短、模糊的描述，譬如研究休闲时间及其感受。对另一半运用更全面的描述，包括说明有一些访题涉及饮酒、吸毒、性事等。研

究发现，给予受访者不同的调查内容信息与应答率（或拒访率）无关，或与应答质量无关。不过，获得了较多内容信息的受访者在自访问卷中被问到对相关访题的感受时，沮丧感和尴尬感更少，也更愿意报告他们实际有的行为。

对知情同意的理解

Singer（1978）研究了知情同意的内容对调查误差的影响。

研究设计：A2×3×3因素设计，区分了（1）研究描述的长度和细节（详细的和长的，与模糊的和短的）；（2）保密承诺的程度（无、保证、绝对保证）；（3）签署知情同意（不签署、事后签署、事前签署）。大约2 000位成年人被随机分配到18个实验组，每个访员都要管理18个组的访问。访题的内容涉及休闲活动、精神健康、饮酒、抽大麻以及性行为。

研究发现：在三个因素中，只有签署知情同意因素影响了样本无应答（不要求签署知情同意的有71%的应答率，另外两组的应答率为65%～66%）。大约7%的受访者因要求签署知情同意而拒访，尽管他们原本愿意接受访问。已有保密承诺对访题无应答有影响，高保密承诺的，则敏感题的访题无应答少。不过，事前签署知情同意的，对敏感行为的报告则更少。

研究局限：由于每个访员要管理18个组的访问，在样本分配上难免出错。由于没有外部效度数据，对应答质量的分析仅依据未经检验的假设。

研究意义：保密承诺对敏感数据缺损的影响说明，在这类测量上，承诺或许比调查测量更重要。有人愿意应答却不愿意签署知情同意，说明知情同意的内容与无应答率有关。

运用同样问卷的一项研究发现（Singer and Frankel，1982），使用模糊句子说明调查目的对应答率和质量没有影响。追随这项研究的主旨，询问了受访者的观点，即是否应该告诉参与者研究的目的，即使被告知的信息有可能改变受访者的应答。其中，让一半的样本仅了解告知的困境，让另一半的样本还知道政策和预算相关的信息。不可改变的是，两组样本都说，不管研究结果是什么，研究者都应该事前告知研究内容。当只涉及内容时，68%的人选择了事前告知，25%的人选择了可以事后告知；当涉及酬劳时，65%的人选择了事前告知，29%的人选择了事后告知，每组之后5%的人说不用告诉潜在的参与者研究目的，每组1%～2%的人说他们根本就不该做研究。这些偏好与大多数研究者的实践一致，为了避免应答偏差，通常粗略、简短地说明研究目的。

在1978年和1982年的研究中，除了一些个人特征以外，只有四分之一的样本说，调查中的有些问题，研究机构“根本就不该问”。这里，不同的实验环境获得了不同的应答。与对受访者做了简短介绍的比较，详细知道了调查内容的受访者很少说研究者根本就不该问，不过，与更少知道方法目的的受访者比较，更多知道方法目的的受访者则更多地说研究者根本就不该问。对四项实验条件下的应答分布进行观察表明，既不对敏感内容做说明，也不对方法目的做说明，就会有负面后果。这就意味着，如果不事先说明，受访者出于方法目的很难判断为什么要问敏感问题。对那些说研究者根本就不该问这些问题的

受访者，冒犯了他们的正式性、收入等经常被问到的问题（参见表11.5）。

表11.5　受访者说“根本就不该问”的敏感访题类型，自认为被冒犯的比例

访　题	有冒犯的百分比
涉及性的访题	19.2%
涉及收入的访题	9.4%
涉及饮酒的访题	5.1%
涉及大麻的访题	4.2%
涉及精神和身体健康的访题	1.2%

数据来源：Singer，1984.

尽管这类研究设计有瑕疵，如由访员展示访题，而不是自访，结果却说明了受访者对不同调查访题的敏感性以及提前告知调查议题的影响。

正如前面指出的那样，尽管调查实验发现几乎没有证据表明告知调查内容和目的的多少会影响调查参与率，实验室实验还是发现了两个变量更强的影响。例如，Berscheid，Baron，Dermer和Libman（1973）在心理实验中通过逐步给被试更多的既有结果信息，譬如Asch（1956）、Milgram（1963），他们发现，被试知道得越多，就越不愿意作为被试。Gardner（1978）告诉被试，如果噪声干扰太大，就可以中断任务。这一告知与没有告知比较，似乎消除了噪声对操作的有害效应。King和Confer（引自Horn，1978）告知被试不同的真实目的，希望看看使用“我”和“我们”是否会带来不同的回报。对被告知了准确、完整信息的被试，假设实验效应便消失了。

这些实验室结果显示，告诉被试其即将参与实验的更多信息会影响到其参与意愿以及研究结果。至少在某些情况下，调查研究者的两

难的确是：是冒着参与率降低（最终会影响到结果）的风险告诉潜在的参与者更多信息呢？还是违背伦理准则获取知情同意以保证应答率进而避免对结果造成不利影响呢？注意，这样的策略并不总是成功的。例如Aronson和Carlsmith（1969）指出，如果不对研究对象说明研究活动足以令其骄傲，他们就会自己建构假设来猜测实验者试图获得什么并照此行动，支持假设或证伪假设。

Couper，Singer，Conrad和Groves（2008）等做过实验，告知受访者应答调查访题可能带来的风险，即信息有可能为其他人所知，譬如姓名、地址。换句话说，违背保密原则。运用情境性，他们描述了不同的风险程度和不同的敏感和不敏感调查类型，他们发现对信息泄露可能性的说明没有影响受访者参与调查的意愿。不过，当告知受访者泄露信息的风险程度较高时，受访者参与调查的意愿会显著下降。

到底为受访者提供多少有关调查的信息才合适，是一个不大容易回答的问题。Smith（1979，p. 15）认为，调查者占据了知识的优势，在涉及此类问题的决策中容易产生偏见，因此，“有理由把这个问题交给评审小组去评判”。我们则认为伦理审查委员会也应该考虑参与者的偏好，尽管不一定采用“社区代表”这种常见方式。由此，我们建议继续进行系统研究不同类型的调查中参与者的规则与偏好。在确定是否支持研究时，这类偏好不一定会纳入考量，不过，也需要认真对待。总体上，过去的研究在这一点上是可靠的。

综上所述，很少有文献证明告知访谈内容信息的多少会影响受访者的合作或数据质量。不过，实验室实验发现，更实质性的是事前告知敏感访题的内容。当调查涉及敏感议题时，对保密性作更多的说明有利于促进受访者合作和数据质量。在这个领域，更值得期待的是对

公共偏好和容忍度的研究。如果没有这类信息，伦理审查委员会的成员就只能凭借主观判断来猜度受访者的偏好了。

研究知情同意在方法研究中的复杂性　Smith（1979，p. 13）发现了伦理责任与通常研究过程之间的冲突，许多实践看起来“对精心设计的、获取可靠知识的能力非常重要”，例如心理实验中的欺骗、隐藏观察，就与任何严格的知情同意不相容。许多调查方法研究的核心议题涉及受访者或访员的刺激反应。例如，在研究访题措辞时就涉及不同程度的诱导（参见第6.6节），以及对访员的监察（参见第9.7节）。在这样的研究中，相应的知情同意又是什么呢？如果让被试充分了解研究意图，则他们就可能依据不同实验刺激而改变自己的行为。不过，研究者更担心的是在测量之前告知受访者会扭曲研究发现。

例子：为了更好地理解1990年人口普查的邮件调查低返回率，美国人口普查局对人口普查参与者进行了一次调查。调查询问了人口、态度等问题，也询问了受访者家里是不是把人口普查问卷寄回了。之后，把调查数据与人口普查文件相关联，把人口普查问卷与调查参与者的应答进行匹配（Couper，Singer，and Kulka，1998）。

在这个例子中，我们没有不做知情同意会有什么后果的信息。例如，一定比例的人口普查参与者调查的受访者，如果寻求其许可并被拒绝，就可能匹配不到他们的人口普查调查表。如果有足够的人拒绝，且他们与允许匹配的人有本质的不同，那么，研究结果就被扭曲了。与此同时，同意匹配者如果知道其调查应答会被拿去与人口普查记录进行比较，则其调查应答可能不同于真实应答，这种情况同样会扭曲研究发现。即使是在调查完成之后寻求受访者对匹配的许可，那些知道自己的应答与人口普查记录不一致的受访者还是会拒绝匹配，

进而让研究者认为两项数据一致的情形高于实际的情形。如果这种效应是实质性的，就得另想他途来既满足伦理需求又使科学的价值最大化。目前，如果伦理审查委员会发现：研究会给受访者带来最小的风险，不会因免除知情同意而给受访者带来危害，而如果没有知情同意就不能继续研究，受访者也会看到完整的研究结论，则会免除知情同意。

研究书面或口头知情同意 到目前为止，我们一直讨论的都是让研究参与者从受访者那里获得知情同意的伦理责任。那么，这种责任要延伸至获得书面签署的知情同意书吗？

许多研究记录了要求参与者书面签署知情同意书带来的有害后果。Singer（1978）的研究应该是这类研究中最早的，结果是，在一项全国性的面访调查中，因要求书面签署知情同意而减少了应答率7个百分点。事实上，大多数拒绝签署书面知情同意的受访者都愿意参与调查，而要求他们签署，他们就拒绝了。在调查前还是调查后要求签署书面知情同意书，对应答率倒是没有影响。Trice（1978）报告了同样的结果，他发现要求书面签署的受访者应答率低于不要求书面签署的。Singer（2003）最近的实验发现，有13%的受访者说愿意参加调查却不愿意因参与调查而签署知情同意书。

鉴于要求受访者书面签署知情同意书对应答率带来的有害影响，我们认为永远不应该要求书面签署知情同意书。签署书面知情同意书保护的不是受访者，而是调查机构。取而代之的应该是有用的备选方案。例如，让受访者签个字，说明已经告知过他们知情同意，并让其留一份备用。应该要求调查机构督促此事，就像其督促是否真的进行过调查一样。只有依共同准则要求书面签署知情同意时，才签署书面知情同意。放松对书面签署知情同意的要求并不意味着放松对受访者

面对风险和要求自愿参与的告知。相反，当参与调查会带来风险高于最小风险 （minimal risk）时，更要提醒去告知受访者足够的信息，确认他们理解了他们所知道的，以及确认他们有足够的时间进行判断并决定是否参与调查。如果面对的是最小风险或没有风险，就是另一种情形了。在这种情形下，受访者的应答或许就是最好的证据，证明受访者已经知情同意，且自愿参与。

下面是一个在面对最小风险时调查说明的实例，消费者调查（SOC）是密歇根大学做的。“我是从密歇根安娜堡的密歇根大学在给您打电话。我们大学正在进行一项全国范围的调查项目。我们希望了解人们对国家经济的想法和感受，为此，我们希望对您进行访问。我希望现在是与您访谈的合适时机。”在调查开始之前，要求访员读出“与您的访问是保密的，且完全是自愿的。如果我们提的问题中，有任何您不想回答的问题，您告诉我就是了，我们就会接着问下一个问题。”如果受访者了解这些，且同意继续访谈，则知情同意就完成了。

调查中的知情同意小结 调查研究者通常让访员口头介绍或者在自访问卷中以书面形式告知知情同意的相关规则。到今天为止的方法研究说明，知情同意的内容对受访者的合作意愿与数据质量只有轻微影响。与此同时，在介绍调查时说明访题具有敏感性，可以缓解访题内容给受访者带来的尴尬和沮丧。不过，如前所述，对知情同意的影响需要更多、更好的研究。

不断有证据表明，如果要求受访者签署知情同意书，有些人可能愿意提供应答却因需要签署知情同意书而拒访。伦理审查委员会应该免去书面签署知情同意书的要求，如果符合共同准则的话。对可能带

来更多风险的调查，研究者需要做的是尽量减少因签署知情同意书而给调查合作带来的损害。

11.7.2　研究保密和调查参与

正如已经提到过的（第11.5节），突破保密性是最可能给受访者带来危害的途径。许多调查都会涉及敏感话题（譬如收入或饮酒），甚至涉及污名或违法行为。此类信息的泄露，就会让受访者的声誉受损、丢掉工作，甚至面临民事或刑事惩罚。此外，至少还有两个理由让我们认真对待保密性问题。第一，违反保密性原则违背了对参与研究者尊重，即使泄露的信息不会造成任何社会、经济、法律或其他危害（参见Citro，Ilgen，and Marrett，2003，Chapter 5）。第二，违反保密原则会伤害调查查自身，因为涉及隐私、保密的信息正是潜在受访者拒绝参与调查的理由。第11.7.2节说明的正是这一点。尽管经验研究的数量还不是很大，其中许多还是美国人口普查局做的，从这些研究进行推广也有局限性，不过，所有这些都说明即使调查合作商有小的负面影响，统计上也是显著的。

早期的研究（National Research Council，1979）试图探讨改变人口普查中承诺的保密时间长度如何影响了邮寄普查问卷的返回率。在研究中，对受访者采用不同的保密承诺说明方式。随机样本第5个收到的保密承诺是“永远”，另一个是75年，第3个是25年，第4个没有说明保密时间。第5组被告知，他们的应答可能会给其他机构或公众。收到了这些说明并阅读了的，包括实验处理，拒访率从1.8%到2.8%不等，保密承诺强的，拒访率低；保密承诺弱的，譬如可能与其他机构

甚至公众分享数据的，拒访率高。百分比的变化虽然不大，统计上却是显著的。

有两项研究考察了人口普查的返回率和态度调查中对隐私与保密的表述。第一项研究通过匹配人口普查的应答，于1990年夏天即十年一度的人口普查之后几个月进行面访态度调查。（调查结果显示）隐私关注指数和保密关注指数都能很好地预测普查邮递问卷的返回率，也很好地解释了人口特征净方差的1.3%（Singer，Mathiowetz，and Couper，1993）。2000年人口普查时，这项研究又重复做了一次，用盖洛普随机数字调查访问了2 000户，并与其普查应答进行匹配（Singer，Van Hoewyk，and Neugebauer，2003）。对隐私和保密的态度能够估计到普查邮件返回率方差的1.2%，与1990年的结果几乎一致。对一人户的家户，相关性更强，即返回普查问卷的行为与调查中显现的态度之间，相比整体而言，具有更高的一致性。相信普查数据会因司法目的而误用的，正如用一个三访题的指标测量的，对普查问卷的返回率具有显著的负面影响。

另外两项研究也支持了顾虑隐私会降低人口普查应答率的揭露。Martin（2006，Table 4）的研究表明，收到长表或顾虑隐私的受访者，更有可能返回不完整应答的问卷或干脆不返回。Junn（2001）的实验也显示，在询问用于考察人口普查隐私顾虑的访题时，与聆听过说明的或对照组的受访者比较，实验组的受访者不大可能应答长表访题。

Bates，Dahlhamer和Singer（2008）的研究分析了国家健康调查（National Health Interview Study）访员编码中由样本家户表述的“台阶顾虑”，表明对隐私与保密的顾虑显著地预测了暂时性拒访而非最终拒访，并建议，如果访员能够成功地强调这类考量，受访者不

一定会拒访。鉴于这些数据来自于面访调查，或许在不同的调查模式中，因隐私顾虑而导致最终拒访的情形也不相同，部分情况可能是因为访员没能在电话中有效地说服受访者。（对美国人时间利用调查［American Time Use Survey，ATUS］的小量受访者和非受访者的应答分析调查［Response Analysis Survey，RAS］采用了电话调查模式，与受访者的保密性顾虑只有24%比较，非受访者的顾虑更多，达到了32%。由于ATUS非受访者对RAS的应答率非常低（32%），而受访者的应答率高达93%，因此，受访者与非受访者之间的区别也许更大［O'Neill and Sincavage，2004］）。

涉及隐私/保密性顾虑与调查参与实验的最后一部分涉及询问社会保障号（SSN）的影响。1992年人口普查局的实验表明，在邮寄问卷调查中，询问受访者的社会保障号会让应答率下降3.4个百分点，且增加差不多17个百分点的访题无应答（Dillman，Sinclair，and Clark，1993）。在关联2000年人口普查的实验中，如果询问家户所有成员的SSN，邮寄问卷的应答率在高覆盖区域会下降2.1个百分点，应答率为所有地址的81%；在低覆盖区域，如大量的乡村黑人、西班牙裔人口，以及租住人口，应答率则下降2.7%。如果只询问第一个受访者的SSN，则SSN缺损值为15.5%，如果接着问家户中第2～6人的SSN，缺损值会逐步增加（Guarino，Hill，and Woltman，2001，Table 5）。

实验还讨论了到目前为止正在进行的调查或纳入与普查应答配对的调查，这些调查都有很好的外部效度，只是能提供的额外信息非常有限。其他的一系列研究，许多是实验室实验，我们知道，如果访题敏感，越是涉及性行为、药物滥用、其他污名行为、财务信息，就越需要更强的保密说明才能有较高的应答率或应答质量（Berman，McCombs，and Boruch，1977；Singer，Von Thurn，and Miller，

1995）。另一方面，如果研究议题无伤大雅，越是强调保密性，则越可能产生相反的结果，导致更低的参与意愿、更低的参与，以及越是怀疑对信息的利用里会有点什么（Frey，1986；Singer，Hippler，and Schwarz，1992；Singer，Von Thurn，and Miller，1995）。

11.8　保密的管理与技术过程

违反保密原则的事，可能发生在多种情境下（National Research Council，2005；Ch. 4）。这里，我们考虑两种调查组织可以控制的情形：粗心和统计泄露。

11.8.1　行政过程

为防止因粗心而违反保密原则，许多调查机构都有书面的保密承诺要求员工签署。例如社会研究所的保密承诺，如图11.1所示。要求承诺保护受访者的隐私和保密信息，并说明在聘用期间一定要遵守承诺。每个员工每年都要续签保密承诺。

密歇根大学

社会研究所

保护受访者隐私承诺

我已经阅读社会研究所保护受访者隐私的政策并承诺将遵循。特别是：

除非是与受访者参与的研究直接有关的研究人员，我将不泄露任何受访者（或家庭成员、知情人）的姓名、地址、电话号码或其他可识别信息给任何个人。

除非是与受访者参与的研究直接有关的研究人员，且得到项目主观或授权代表授权，我将不泄露可识别受访者或知情人的应答内容或实情给任何人。

除非得到与受访者参与的研究直接有关的人员授权，我将不与任何受访者（或家庭成员、雇主，或其他与受访者、知情人有关联的人）联系。

除非依据社会研究所以及我所在中心的政策或程序，我将不泄露数据集（包括非严格意义上的公共用途或其他严格意义上的用途）。

我同意服从此承诺及相关政策：(1)作为我供职（如果我是社会研究所的雇员）的条件，和/或(2)继续合作和与社会研究所联合（如果我不是社会研究所的雇员，如学生、访问学者、外部项目主管或调查主管等）的条件。

如果我督导的非社会研究所雇员接触到了社会研究所受访者的数据（不是非严格意义上的公开数据集），我保证让那些雇员谨循本承诺书和相关政策的标准以保护社会研究所受访者的隐私、匿名性以及保密性。

签名：______________

印刷体姓名：__________　　　　日期：__________

图11.1　由研究团队成员对受访者隐私作出的承诺

在美国政府机构，对承诺的强制还包括了额外的法律处罚。例如，在美国联邦统计调查部门工作的人员如果违反了保密承诺，将会面临高达5年的监禁和高达250 000美元的罚款。

遵循对受访者保密承诺的另一个必要组成部分是，工作场所的承诺规则非常严厉。例如，密歇根大学社会研究所在机构层面有一个保密性和数据保密委员会。1999年4月，委员会发布了一个文件，“保护敏感数据：研究人员需要遵守的原则与实践”，包括了14条针对纸版和电子文档的原则（参见表11.6）。这些原则说明，对受访者保密信息最大威胁或许来自粗心，而不是有意。

表11.6　保护敏感性数据的原则与实践

1. 评估风险。对包含识别信息的材料要特别谨慎;包含清理过的以及累计数据的文档,可以作为,也可以不作为保密文档。
2. 评估自己管辖的所有数据的敏感性。
3. 采用适宜的安全措施。删除直接识别信息,如姓名、地址、社会保障号。含有敏感信息的问卷或磁带,如药物滥用、医疗状况,应保存在上锁的柜子中。有时,含有开放题的问卷也可能泄露受访者或他人的识别信息。
4. 不要在自访问卷中纳入可识别个人的信息。如果要搜集个人信息,需专门提供单独的信封。
5. 对包含受访者个人信息的封面页,要保存在上锁的柜子中。
6. 要像对待纸版信息那样,在物理上保证电子文档的安全。
7. 对包含敏感材料的加密硬盘要特别保管。
8. 在硬盘中,将敏感信息与非敏感信息分开保存。
9. 考虑对敏感信息加密。
10. 考虑保密措施的成本收益。
11. 谨记所有电子文档的物理位置。
12. 谨记所有存储系统的备份状态。
13. 应该了解电子邮件在传输中是可被侦测的。
14. 删除文档时要谨慎。大多数被删除的文档是可以被恢复的,除非采用特殊删除方法。

数据来源：ISR Survey Research Center，*Center Survey* ，April 1999，pp. 1，3.

11.8.2 技术过程

对受访者承诺保证其在调查中提供数据的保密性是容易的，要兑现承诺却不那么容易。研究者对受访者承诺要保护数据，且会保留多个报告版本用于分析。不过，分析数据的人并不是直接对受访者承诺的人。的确，现在的一些数据档案馆存储着几千项过去几十年的调查数据。随着可用的管理和其他大型数据的增加。调查研究者需要保证抗拒“数据黑客”用管理数据与调查数据进行匹配，进而识别出受访者并获得受访者的个人信息。

尽管对一些调查数据而言，对试图识别调查受访者的动机需要动用想象力，但对另一些数据而言，则非常简单。以全国药物使用与健

康调查（NSDUH）为例，有时候会问到在家的未成年人，此时就需要对家长告知知情同意，让家长知道孩子参与了调查。同时，家长也应该知道调查的议题。很容易想象的是，很多家长会希望知道孩子有没有对访员报告药物滥用。如果他们有动机知道这些，且有技术操纵计算机文档，就有可能查到孩子的应答记录。这就涉及对受访者个体详细信息的入侵。

“披露限制”指的是评估个体受访者被识别的风险并试图降低风险。这里有两个重要的途径：

1）把数据接触的范围严格限制在做出了保密承诺的人员之内。

2）严格限制调查数据披露的内容。

严格限制数据接触　许多研究项目会同时强调几个问题，以便让不同的研究者从不同的角度获得研究发现。对调查而言，尤其如此，常常会测量受访者的几百个变量属性。

如果数据集含有每位受访者丰富的信息，即使数据集中没有姓名和地址，也有可能识别出受访者。例如，样本中的某个人有独特（或稀有）的特征，就很容易根据调查数据列多维表以识别（例如在大学生数据集中，就很容易识别一位12岁、身体残疾、攻读物理和艺术史双学位的学生）。在数据集中保留这样的数据或许非常重要，可一旦公开数据，就会对保密承诺构成威胁。

在这种情况下，一些调查机构会延伸保密承诺的覆盖范围，即数据分析者或中间机构需要得到授权。在这样的安排下，分析者对保密的承诺与初始调查者就一样了。这样的安排或许会涉及研究机构之间

的法律承诺，也许会要求远程查验数据安全环境。如果远程用户违背了承诺，或许还面临罚金或其他处罚。美国人口普查局采用的是另一种模式，即建立一组研究数据中心，在那里研究者经过特许，且在符合人口普查局保密要求的条件下，远程接触存储在人口普查局的数据文件。分析的所有输出需要经过人口普查局人员的查验，以让泄露风险降至最低。

严格限制调查数据披露的内容　对调查研究者而言，实践中通常的做法是降低非冒险披露的可能性。

1）尽快将姓名、地址、电话号码或其他直接识别信息从受访者数据中分离出来。

2）限制数据文件中地理信息编码的粗细（有时候，地理信息可能是分析变量，如城市名称、背景数据或初级抽样单位），以使受访者数据不可能被识别。

3）对外部人员核查可能引导至识别的数值数据（包括单变量和多变量。如非常高的收入报告值，如果辅之以其他数据，就可能导致泄露）。

如果核查数据时发现存在非冒险泄露的风险，一系列的统计过程在限制调查估计的同时，也可以降低泄露的风险。

在研究泄露限制中，有一个基本的、常出现的概念。第一，**总体唯一**（population unique）指在目标总体中，如果对某些分类变量集采用交叉表，则某个单元格只有一个要素（也就是说，对给定的变量，只有唯一值）。**样本唯一**（sample unique）指的是目标总体的

给定样本（参见Fienberg and Makov，1998）。这是两个相关的概念，相对于要素，在数据集里，采用属性交叉方式更容易检测到总体唯一，因为它非常普遍。如果相对总体规模而言样本规模较小，则可能有许多的样本唯一，且不是总体唯一。

第二，泄露风险 （risk of disclosure）与泄露导致伤害（harm from disclosure）之间的区别（Lambert，1993）。泄露风险指用发布的数据识别出个体的概率。泄露导致伤害指泄露数据中的什么信息给个体带来的后果。Lambert注意到如果入侵者确认已经识别，他们会以同样的方式确认识别是否正确。不幸的是，目标个体也会因入侵者披露虚假信息而受到伤害，其受到的伤害与披露真实信息是一样的。

当不变的数据记录指向不可容忍的再识别风险时（研究者定义为“不可容忍的”），就可以在发布数据之前运用多种方法进行数据变化。包括：

1）地理阈值。

2）数据变换。

3）再编码。

4）添加噪声（扰乱）。

5）补值方法。

地理阈值（geographic thresholds）指发布数据中最小的地理区域规模。现在，美国人口普查局对总人口小于10万的区域不发布可识别地理信息（Hawala，2000；Zayatz，2007）。一个样本的精准位置信息是最有利于识别的，对调查研究者而言，限制详细地理位置信息是重要的考量。

数据变换（data swapping）是对整个数据记录的报告数据值进行交换（Fienberg，Steele，and Makov，1996）。数据变换允许调查组织诚实地说明入侵者无法保证找到的数据会准确定位到个体。真正的不确定性来自在数据变换中纳入数据项的百分比和样本值的百分比。不过，如果研究者不说明变换率，则察觉到的不确定性会更高。当然，在数据变换中的挑战是，在经过数据变换之后的统计计算值与之前的有多接近。

再编码（recoding）就是改变离群值的值。含有这些值的样本，极有可能是样本唯一的案例。再编码就是把它们放回到与其他案例分享的类别中去。例如，“顶端编码”（top coding）就是把变量的最大值在数据记录中赋值的技术（例如，把所有年收入为25万美元或更高的作为一类）。这是对可能是唯一的案例数进行的直接干预，由于裁剪了尾部，也会对统计的某个类别造成信息损失。其他的统计（例如，收入少于10万美元的百分比）则不会受此累计的影响。其他的编码方法，包括把某些变量的值，如年龄、收入归入事前确定的类值。例如，美国人口普查局把8～9.99美元统统归入最近的10美元。

添加噪声（perturbation methods）是运用统计模型对个体数据值进行变换。例如，用一个随机取得的变量值加到某几个访题的数据值上。如果公布的是添加了噪声的数据，则入侵者就不可能知道公开使用的数据与真实的数据之间有怎样的关系，也不可能定位到个体。

如果变量的平均值是0.0，即使均值的总体方差会有所增加（由于添加的噪声是一个随机值），样本均值还是一样的。如果变量之间的协方差不受噪声影响，则可以运用联合噪声，让其协方差一致。涉及的变量越多，方法也就越复杂。一般而言，给变量添加的噪声越大，对数据的保护也越好，不过，损失的信息也越多。

补值方法（imputation methods）就是采用补值程序，用另一个值替代受访者的报告值。为了避免数据泄露，Rubin（1993）第一次提出了这种大胆的想法，即建立一个完全合成的数据集。Fienberg（1994）也提出了同样的想法。这个完全构造的数据集不包含任何受访者的任何真实报告值。在某些情况下，这种方法为数据提供了完全的保护。如此，补值模型就承担了让合成数据集的主要统计属性的统计值与原始数据的一致的负担。Rubin建议同时产生多个补值数据集，如此，可以在经验上评估补值方差。还有人建议在数据集中只对敏感数据或可识别的敏感数据采用完全补值方法，用以保护数据（Little，1993）。

在经验应用中，这些想法遭到了批评。例如，Abowd和Woodcock（2001）指出，用顺序回归补值产生的合成数据与原始数据有许多相同的统计属性。美国人口普查局现在既用全合成（为所有记录生成合成数据），也用部分合成（仅为部分数据记录合成数据），具备了合成人口和其他数据的能力。用收入与项目参与（SIPP）调查产生的连接数据的合成数据涉及了从社会福利部门和税务部门获利的历时雇员-雇主收入报告，这个数据可以用于研究者进行测试了（Abowd，Stinson，and Benedetto，2006）。

当变量分布极偏时，泄露限制方法就不限于微观数据了，累计数据也一样。例如，用既有调查数据列表中包含的、从所有调查对象贡

献的数据累计出来的数据（如递送值）。不过，有些单元格的数比较小，就揭示了已知企业的属性。正如Felsö，Theeuwes和Wagner（2001）注意到的，有两种方法可以用来判断一个单元格的数据是不是敏感：

（1）（n，k）规则。如果一个小数字（n）受访者贡献了相对于所有单元格而言较大的百分比（k），那么，这个单元格值就是敏感的。

（2）p 百分规则。如果在p 百分数内任何贡献值是可以估计的，则这个单元格值就是敏感的。

大多数作者在实践中采用的是（n，k）规则。一旦某个单元格被认定为敏感，就可以作多种选择。隐瞒 （suppression）规则意味着从表中对其进行隐瞒。例如，如果少数企业（譬如3家）占据了总销售较大比重（譬如70%），则这个单元格的值就可以隐瞒不报。如果如实公开，则可能依据公开数据识别出是哪些企业（甚至其营业额）。除了隐瞒以外，也可以采用再编码或合并的方法进行数据保护。其他的技术，如为单元格或潜在的微观数据添加噪声，也可以用来保护表格数据。

未来，还会有新的方法用来防止对调查数据的滥用，以让调查数据获得更加广泛的利用。当然，随着可用数据的增加，可入侵数据的资源也在增加。在这个领域需要更多的研究，包括对受访者伤害风险观念，以及如何就此进行更好沟通的研究（Couper，Singer，Conrad，and Groves，2008）。

11.9　小结

把科学活动的社会规则和规范运用到调查方法的有：对伪造数据、报告作弊以及剽窃他人成果进行惩罚。这些惩罚还会延伸至支持研究的工作人员，包括调查访员。在数据搜集中，访员作弊的风险小，却始终有，特别是在面访中。不过，通过科学伦理培训、观察访员的工作、回访受访者，以及对搜集到的数据核查，发现作弊率在下降（参见第9.8节 ）。

调查的专业组织，如AAPOR，制订了一些伦理规则用以规范调查方法学者，包括客户、公众，甚至受访者的行为。对规则最为系统的解释适用于如何对待受访者。其中有三项原则：善行、公正、对人的尊重。善行原则旨在激励人们避免伤害任何调查受访者。公正原则在于平等分摊参与调查的负担，在概率抽样调查中常常如此。对人的尊重原则提出了给予样本对象在获知调查内容以及潜在的风险和收益之后拒绝参与的权利。在美国许多学术和政府研究中，伦理审查委员委员统管着在以人为研究对象的研究中这些原则的执行。

依据通用规则，对知情同意原则（参见表11.4）还有特殊要求。调查中的知情同意程序通常涉及用语言说明调查内容的特征以及对数据保密的程序。告知多少调查内容对受访者的合作和数据质量似乎只有微弱的影响，尽管实验室实验获得的影响较大。要求受访者书面签署知情同意书会让原本愿意应答却不愿意签署的受访者拒访。根据通用规则，鉴于要求书面签署知情同意书会损害调查结果的质量，对只有最小伤害风险的调查而言，不应该要求书面签署。

保护受访者免受伤害，善行原则的导向是：支持对受访者承诺保密的程序。这些程序包括让职员接受培训，并正式承诺，在联邦机构，对违反保密承诺的给予法律严惩。包括运用各种方法把保密承诺延伸至原本不是研究队伍成员的数据使用者，要求其不可以用分析数据识别受访者。最后，还包括用不断增加统计方法限制运用公开的调查数据文件识别调查受访者数据记录。具有唯一或近似唯一多变量值组合的数据有着更高的泄露风险。减少这类数据泄露风险的技术包括数据变换、再编码、增加噪声以及补值。如果运用调查数据建构交叉表对样本单位（通常是商业机构）有识别风险时，通常会采用单元格隐瞒方法。

调查研究者必须主动遵守上述伦理准则。对没有经过训练的人，这些准则是不会自动钻到头脑中来的。对调查队伍的新人而言，通常要接受培训，以了解以人为研究对象时涉及的基本原则和处理方法。美国国家健康研究所（NIH）有一个基于互联网的培训模块，主要涉及生物医学研究的研究者培训（参见http://www.nihtraining.comJohsrsite/）。最后，保护受访者的权利还依赖于调查研究职员的参与。正因为如此，成功地构建一个强规则工作环境，强调知情同意的重要性，避免伤害受访者，以及遵守保密承诺是践行伦理原则的最有效途径。监管只能是这类内部规则的补充。

在以人为对象的调查研究中涉及的伦理实践也在不断演化，不仅仅是因为社会对滥用的顾虑，也因为调查方法学者不断在提供新的工具以减少研究中的误操作，践行保密承诺，以及改善调查中的知情同意过程。

关键词

剽窃（plagiarism）

造假（fabrication）

善行（beneficence）

对人的尊重（respect for persons）

保密证（certificate of confidentiality）

统计披露（statistical disclosure）

总体唯一（population unique）

泄露风险（risk of disclosure）

地理阈值（geographic thresholds）

再编码（recoding）

补值方法（imputation methods）

作弊（falsification）

伦理审查委员会（Institutional Review Boards，IRBs）

公正（justice）

知情同意（informed consent）

保密性（confidentiality）

最小风险（minimal risk）

样本唯一（sample unique）

泄露导致伤害（harm from disclosure）

数据变换（data swapping）

添加噪声（perturbation methods）

隐瞒（suppression）

进一步阅读资料

Citro, C., Ilgen, D., and Marrett, C. (2003), *Protecting Participants and Facilitating Social and Behavioral Sciences Research* , Washington, DC: National Academy Press.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001), *Confidentiality* , *Disclosure* , *and Data Access* : *Theory and Practical Applications for Statistical Agencies* , Amsterdam: North-Holland.

作业

---

1. 美国许多大学都为研究者在以人为研究对象的伦理处理上提供基于互联网的训练。如果你们学校也有类似的训练模块，完成它。如果你们学校没有，登录国家健康研究所的网站（http://www.nihtraining.com/ohsrsite/），完成它。

2. 在下列情形下，知情同意提出了什么议题？如果你是伦理审查委员会的成员，你会批准研究活动，还是要求有所修改或增加？判断对每个情形的回答。

   （a）作为面访的一部分，要求访员观察在受访者的客厅有多少看得见的书。

   （b）在调访中，作为方法研究的一部分，录下访员与潜在受访者的初始互动，以供以后分析。

   （c）在征求家长意见把与孩子的访谈作为儿童互动研究的一部分之前，访员对父母与孩子在公园的互动做结构式观察。观察用于样本选择，也作为每位受访者数据搜集的一部分。

   （d）作为研究投票是否受调查资助者影响的一部分，告诉有些受访者资助者是不同于组织调查的另一个组织。

3. 研究者希望确认女性流产报告到底有多真实。从某个医疗门诊的医疗记录中，获得了流产女性的姓名和住址。接着，给这些女性发送调查参与的邀请。除了一些事情以外，邀请信还说：

   “调查的目的是为女性的健康和影响健康的因素搜集数据，还有女性对提供医疗服务的机构的满意度，以及如果有机会重新组织时女性的选择。研究最核心的部分是直接从人们那里搜集数

据的调查。受访者，包括您自己，都是采用随机方法从注册地址中抽选的……”

（a）如果有的话，您认为这项研究违反了什么伦理原则？

（b）您认为用这样的程序搜寻受访者会有怎样的后果？

（c）您认为还有不同的做法吗？怎样做？

（d）如果您是受访者，且知道了自己的姓名是如何进入样本的，您的反应会是什么？

（e）假设这是研究人们自报投票行为的准确性，样本的姓名来自选举登记列表，且不告诉受访者他们是如何进入样本的。您认为这项研究违背了任何伦理原则吗？作为受访者，如果您发现了样本是如何抽出来的，您的感受如何？

4. 作为课程作业的一部分，作为社会心理学导论的本科生，学生们被送往纽约市的不同区域，询问1 520名路人一个简单的问题，即寻求不同帮助，且用不同的方式寻求。对求助的不同应答，构成了对利他承诺普遍性及其影响因素的初步回答。

（a）如果有的话，您认为这类研究违反了什么伦理原则？

（b）这类研究的优点有哪些？

（c）这类研究还有其他做法吗？如果有，怎么做？

（d）如果您在以人为对象的委员会，您会批准这类研究吗？为什么批准或不批准？

5. 作为对无家可归者面访的一部分，为他们提供100美元的酬劳。伦理审查委员会没有批准这项研究，理由是，酬劳带有强制性。您如何应对伦理审查委员会的决定？

6. “推送式民意调查”指政治竞选中用电话方式在不征求人同意的条件下向许多人提问调查问题，让人们对竞争对手留下坏印象。例如，“如果我告诉您候选人A被证实从事毒品贩卖、谋杀，会改变您的观点，认为他应该当选吗？”这样的实践违背了什么伦理原则？

7. 您完成了一项9～12岁儿童的调查，在调查中，您从孩子父母那里获得了书面签署的知情同意书，且得到了许可与孩子进行单独访谈，即不允许父母旁观访谈。访谈完成之后，父母要求查看孩子的应答。这项要求提出了什么伦理议题？如何避免？

8. 下面是互联网调查中知情同意书的模板。审核模板中知情同意书的每一个部分。您看到了这份模板在遵循知情同意要求中的任何不足吗？

欢迎来到“看护”：抑郁与支持调查（HUM00001234）

密歇根大学心理学系的John Jones博士和Sara Smith博士邀请您参加这样一项研究，即在人们看护严重疾患家人时所经历的抑郁。研究的目的在于设计更好的看护支持项目。我们邀请您参与，是因为您最近参加了密歇根大学的看护支持小组会议。

如果您同意参与研究，就请您填写涉及您看护体验的在线调查问卷。调查大约会进行30～45分钟。

尽管从调查中您不能直接获益，不过我们希望这项研究能改善为他人提供看护的人员的社会支持系统。

您的应答是匿名的，意味着研究者不可能通过调查应答识别您。调查软件也不会搜集您和您电脑的识别信息。对研究的结果，我们打算公布，不过不会包括任何可以识别到您的信息。

参加这样研究是完全自愿的。即使您现在决定参加，在接下来的时间里，您可以随时改变主意，退出。您也可以选择不回答某道访题或跳过任何部分。点击“下一步”按钮，进入下一道题就好。

如果您有任何对研究的疑问，您可以根据下面的联系方式联系密歇根大学的Jones博士。123 East Hall，Ann Arbor，MI 48104，（734）123-4567，jones@umich.edu。

经密歇根大学行为科学伦理审查委员会审核，这项研究获准实施。

点击下面的链接，意味着您同意参与这项调查研究。

www.caregivingsurvey.net

如果您不希望参加，请点击您浏览器工具条上的“X”以退出。

12　调查方法常见问题

12.1　导言

至此，这本书已经展示了调查方法的许多基本原则和实践。所有这些章节描述的都是如何实践这些原则，定义了调查质量的框架。关键术语定义了这个领域的技术性术语。文本框则通过一些经典文章展示了这个领域关键类型的科学探索。最后，每章后面的习题则模拟了调查方法专家在他们工作中强调的问题。

不过，你可能还是有疑问，致使你不能在头脑中把这些知识整合为一个体系。在本书出版之前，读过初稿的人也有这样的感觉。他们的评论帮助了我们对每一章的修订，不过也有一些更一般性的问题，不能放入任何一章。此外，还有一些问题是调查方法专家经常听到的，在这个领域也是基本的问题。

这一章与其他章节在形式上完全不同。采用了常见问题的问答模式。问题用粗体字。你可以浏览本章，寻找自己有兴趣的问题。如果有关联，答案也许会涉及本书其他章节的内容。

12.2　常见问题与答案

这本书描述了这么多误差，调查管用吗？人们能够信任任何调查的结果吗？

《调查方法》一书是一部如何最小化统计误差和最大化调查结果可信度的教材。有鉴于此，本书的关注点就在于为什么会出现误差以及怎么做才能降低误差等科学问题上。每一章都特别讨论了若干影响调查统计质量的因素，以及在调查方法的意义上如何强调这些因素。

在这样的背景下，很容易夸大调查中的误差。当然，有不少调查的确做得很糟糕，根据本书提供的标准，这些调查产生的数据也的确令人生疑。不过，许多调查也采用了极好的流程，调查获得的数据也非常精准。常见的搜集数据方法为概率抽样调查，即从样本获得的数据与样本所在的总体非常一致。调查访题通常经过细致的检测，以消除理解与应答的困难。在大多数情况下，改进过的访题能让受访者一致和准确地应答。访员也接受过集中训练，在大多数情况下，能够按照规程操作调查。

调查方法研究显示，恰当地实施调查可以获得很高的数据质量。例如，1998年CES调查获得职业数与基于实际管理获得数据之间仅有0.2%的差距。NAEP的评估，经过该领域专家的定性和定量研究，获得了很好的评价。在图1.3中，我们展示了SOC的失业期望值与后来实际出现的失业值非常接近。BRFSS成功地侦测到美国肥胖的流行（参见图1.4）。在调查中运用本书的内容可获得非常有用的信息。

另一方法，如果在调查设计和实施中忽略本书的内容，调查有可能获得极有误导的结果。参见国家地理协会的网络调查讨论。

这本书没有提供搜集信息的其他方法，譬如民族志、深度访谈或实验室实验。问卷调查是研究人类的最佳方法吗？

不是的。调查方法是我们的专业领域，也是这本书的关注点。有许多其他方法可用于探讨人类的思想、行为，到底运用什么方法，则取决于要研究的问题。对研究方法的选择还取决于研究的目的，而不是相反。问卷调查产生的是量化估计，不过有些知识不那么容易量化。因此，我们并不建议仅仅采用定量或定性方法搜集信息或尝试理解。的确，在第7.3.2节和第8.3—8.4节具体（用教材）说明了用非问卷方法探讨问卷调查统计属性的调查研究结果。调查方法学者采用多种方法试图说明问卷调查是如何管用的。事实上，我们相信，如果问卷调查方法能够辅之以其他方法会更好。

问卷调查方法并不是搜集信息的僵硬方法。调查方法强在通过统计推论到总体，弱在对影响人类思想和行为的复杂机制获得丰富理解。对后者，则应该采用其他技术。对分享基本语言和文化知识的总体，调查也是很好的方法。而在多样化的总体中，其他的技术也许更合适。

调查要花多少钱？

这个问题有点像“买一幢房子要花多少钱？”。如果买别墅，就要花几百万美元，也有棚户，几万美元就够。每一种，都会有买家。

现在你知道了在调查统计中，有许多方面是需要复杂的努力的。如果抽样框没有准备好（如区域概率样本），研究者首先要建抽样框，要花不少钱。如果要求有较高的统计精度（抽样方差小），就需要大样本，进而也增加抽样和数据搜集费用。如果多种复杂问题碰到了一起，就需要CAI应用或复杂事后清理。如果目标总体没有应答动机或没有能力自访，就需要聘用、培训、督导访员。如果目标总体是抽

样框的一个子集，则还需要识别合格受访者的费用。如果目标总体很难联系，就需要费用用于不断追踪。如果目标总体不愿意，就需要考虑激励或拒访转换，这都需要费用。如果没有这些挑战，那么，调查就会相对便宜。如果全都出现，则希望降低调查中的观察误差和非观察误差就需要很大的成本。除非预设所有这些设计特征，否则就不可能按照你想象的用“最低的成本”来实施，因为运用最低成本，最简单的方法就是忽略一个或多个对获得良好估计的关键威胁。

民意调查的结果有多可信？在线调查的结果有多可信？

调查结果的可信度是调查目的、设计与执行的函数。因此，要了解对调查结果的可信度，就需要了解调查的设计与执行。如果没有设计与执行的信息，就不可能判断统计的结果。这就是为什么AAPOR披露的第11.4节的内容是重要的。

如果目标总体、抽样框、抽样设计、访题评估、数据搜集方法、样本和访题无应答属性以及数据清理都有记录，那么，你就可以形成自己的判断了。运用本书的知识，就可以进行判断。

没有调查文档也许是对调查统计进行判断最常见的障碍。本书前面的章节已经说明，对同一项调查，运用不同的统计方法可以获得不同的覆盖面、无应答以及测量误差等参数，知道这一点也是非常重要的。因此，没有所谓的“好调查”或“坏调查”，只有好的调查统计和坏调查统计。

调查中，最重要的误差来源是什么？一项精准调查的特征又是什么？如何能快速确认一项调查有多好？

这些问题可能有多重含义：

1）对调查估计质量而言，什么样的误差来源有最大的潜在危害？

对这个问题，没有确定的回答。一项调查统计可能因覆盖性误差、抽样误差、无应答误差或测量误差而毁掉。

2）对调查估计质量而言，什么样的误差来源通常是有害的？

通常的错误包括自选样本、访题措辞不准确以及不恰当的搜集数据方式。自选样本，如那些应答记者问题的人或打900电话的人，这些人明显与整个人群的特征有所不同（再次，NGS就是一个例子）。更准确地说，邮寄问卷常常因低应答率而会出现这样的情况。典型的情况是，在大多数人不应答的情况下，应答的人与调查相关的目标总体作为一个整体的人群是有区别的（例如Fowler，Gallagher，Stringfellow，Zaslavsky，Thompson，and Cleary，2002）。

如果访题不好理解，也会扭曲调查估计。在态度测量中，设定情境或运用结构，都会让应答有偏。

在询问潜在的敏感性内容时，正如在第5.3.5讨论的模式研究一样，采用合适的方法是重要的。与有访员的调查比较，ACASI获得了较高的药物滥用和污名化性行为的报告率。

这本书的一个关键主题是，对所有误差来源要综合起来考量。到底哪种误差是最大的，取决于目标总体、可用的数据和资源，以及调查的目的。

没有抽样框，如何从目标总体中抽样？

我们用一个例子说明，在美国，如果没有基于个人的抽样框，如何在目标总体中抽样：运用区域概率抽样。

有许多目标总体都没有最新的要素列表。区域抽样技术就是能够连接目标总体进行枚举的第一解决方案。假设研究者要抽取学生，却没有学生名单的列表，那么，第一步，先抽取学校，然后到样本学校建构学生列表。零售店的客户就可以依具体商店的不同入口处列表抽取可能的时间段在店里停留的客户，如此，就可以构造一个不同时间进不同入口的系统抽取个体的方案。同样，也可以通过划定海滩区域来抽取海滩用户，构造时间段抽样方案，然后抽取在样本区域、样本时间的个体。

要抽多大样本？最小样本量是多少？

在调查成本许可的范围内，抽样设计应该准许针对关键调查统计量说明可容忍的不确定性。这句话似乎没有回答问题，不过，让我们来看看。第一，样本量只是抽样设计的一个方面，其他的还有分层、整群以及抽选的概率设计。所有这四个方面都会影响统计的标准误，因此，这四个方面需要综合考虑（参见第4章 ）。第二，在调查基础上，做决策时需要考虑到标准误。如果一个统计量的真值是“X”，如果调查估计值为“0.5X”“0.8X”或“0.9X”，研究者会得出同样的结论吗？对访题的应答，要说明标准误是什么（即显著性是“0.5X”还是“0.9X”）。第三，抽样设计一定要依据关键分析目标。大多数调查都是为了获得数量化的估计，不只是一个，且对不同的统计估计值有不同精度要求。此外，估计值通常会针对总体小的子集，而不是整个总体。如果不同的分析目的产生了不同的、理想的样本设计，那么针对总的目标就要作出妥协。第四，把符合分析目的要求的抽样设

计放到调查预算中，有可能因预算过紧而无法实施。此时，就不得不进一步作出妥协。

在回答了要做什么之后，这里还有不做什么。不要使用其他调查使用的样本量，即使过程相同（例如，盖洛普使用1 000个样本，并不意味着这个量也适合你的调查）。不要在你的总体中使用其他总体的样本比例（也就是说，记住第4.3节的内容，样本量的大小几乎独立于抽样框）。

非概率抽样的确不好的证据是什么？

如果你读过第4.2节，就应该注意到概率抽样的两个好处：

1）重要样本统计量的无偏。

2）用一个样本设计来估计抽样方差（标准误）。

把无偏性运用于抽样偏差，就是指没有覆盖性偏差、无应答偏差和测量误差偏差。概率抽样的一个优势在于，样本依据某些统计规律抽选，而不是基于志愿或可用抽选。某些，并不是全部，非概率抽样会依据其中的一个或多个特征，进而增加了偏差抽样的机会。不过，由于在实践中所有调查都会涉及那些偏差（一般是未知水平），概率抽样的第一个特征是，其偏差值不是最高的值。其更大的价值在于，概率抽样可以确定的估计标准误，即反映了重复样本结果的潜在变异性。

这并不意味着概率抽样的每一个个体样本较之相似的非概率抽样的每一个个体样本更好。如果设计者完全了解框总体的属性及其与关键调查统计量之间的关系，并能成功地平衡这些属性在样本中的分布，那么非概率设计产生的样本统计量就会更加接近于总体。如此抽样，运用了不同子集的“配额”（例如，年龄、性别、种族）也很常见。如果统计量与配额有不同的关联甚至不关联，就会产生相当大的偏差。因此，非概率抽样，如果要优于概率抽样，就需要了解目标总体的多个变量的联合分布。通常，我们是不了解的。换句话说，非概率抽样或许会产生概率抽样类似的样本，不过，人们不能保证的是，什么时候非概率抽样管用，什么时候不管用。

我听说在线的志愿者追踪调查获得了与概率抽样调查同样的结果，是真的吗？

志愿者在线（或接触）追踪调查包括了某人在某个时点所愿意参加某种在线调查的电子邮件。追踪调查责任方给抽选出来的受访者发送电邮，邀请其参与，其中包含了链接到在线问卷的地址。故在考察志愿者在线调查的估计值时有3点需要考虑：①追踪群体是如何获取的。②对其参与给予怎样的回报。③受访对象的经验累积是什么。

受访者集代表整个成年人总体的能力是针对调查变量具有不同值的人被纳入受访群体的函数。在概率抽样中，目标总体的所有个体其被纳入追踪群体的机会是已知的。而在在线追踪中，是受访者自己参与的（因此，有可能是对调查很有兴趣而参与的），换句话说，应答就会获得回报。如果是为着回报而参加应答，则其应答是否是经过思考的，就令人生疑。

还有，为参与提供回报或许会产生这样的情况，即一小撮人为了回报同时参与了多项追踪调查。在美国，受访者报告说平均参加了5～8项在线追踪调查是非常普遍的。有些研究显示，一群非常小的互联网用户代表着很大的追踪调查应答者群体。对某些估计而言，这种招募、回报、应答行为产生的结果非常具有误导性。例如，长期追踪的受访者与短期追踪的受访者比较在某些问题上的应答似乎并不相同。出于其他目的，追踪调查产生的结果与其他技术产生的结果似乎相同。只是到目前，对研究方法的研究还没有识别导致偏差的机制。没有对在线追踪调查志愿者参与动机的理解，研究者只能通过与经验比较，进而获得更偏理论的结论。

民意测验专家报告说，他们的结果有一个“加减 X 百分点”的边际误差。这是什么意思？了解结果到底有多精确有什么帮助？

一项调查在大众媒体中报告的边际误差通常指在95%置信区间，以简单随机抽样为基准，估计百分比等于50%的情形（例如，如果50%的受访者报告说2008年的大选，他们准本投给奥巴马）。例如，如果完访数为1 000，则边际误差为：

$$2\sqrt{(0.5)(0.5)/1\ 000}=0.031\ 6$$

或3个百分点。50%，具有典型性，因为它最大化了在分子为抽样方差表达式中的“p（1-p）”。因此，所有百分比都应该在95%的置信区间中小于或等于这个数。

如果不是简单随机抽样设计（例如，涉及不等概率或整群），则需要对边际误差进行审核。每个偏离简单随机抽样的抽样都会增加标

准误，因此，都会低估边际误差。此外，基于样本的子样本计算的百分比，应该有更大的边际误差。计算仅关联百分比统计量，没有除百分比以外的均值、均值差甚至总的信息。最后，边际误差中一般情况下没有包含变量的覆盖性误差、无应答误差或测量误差。在边际误差中，也没有覆盖性、抽样性、无应答或测量误差的偏差。因此，非常重要的是让读者明白，报告的“边际误差”仅仅反映了可能影响估计的总误差的一项误差。

我们的统计教材中没有提到反映整群设计的加权或抽样方差计算，为什么？

我们知道，本书的公式与大多数应用或数理统计导论教材的有所不同。那些教材中的许多分析性统计都假设获得样本的过程是一种抽样取代另一种或类似，如此，统计归纳的观察值是不断持续的、稳定过程的一个数集。因此，观察值集都有特征属性且相互独立。不过，在我们的教材中，一个观察值集的产生过程要更复杂，涉及了分层、整群、不等概率的抽选。如此，本教材的统计量是多种抽样设计的平均值，即是分层、整群以及概率抽样的函数。

在给定抽样方法的前提下，你在统计课堂和本书中学到的都是对的。大多数调查的抽样设计都会涉及分层、整群、不等概率抽样。因此，本书反映的是这个特征。

对一项问卷调查而言，什么是最好的搜集数据模式？

在具体时间、具体地点，为着具体目的，都可能是最好的模式。选择搜集数据的模式是众多要做的决策之一，类似于确定样本量，那

需要分析可能影响估计质量和成本的议题。正如在第5章讨论过的，在某些情境下选择某些模式可能是不明智的。例如，如果目标样本中明显有部分人不接触互联网或没有使用互联网的训练，那么用互联网方式或唯一方式搜集数据，显然不是合适的选择。如果调查旨在搜集非常敏感或个人的数据，采用有或没有访员的自访，看起来就是一个好主意。总体特征、内容、可用联系信息的特征、工作人员能力的可用性以及资源，对任何调查而言，都可能影响到什么是最好的方式。第5章详细讨论了所有模式相关的优秀调查，加上不同方式的组合。调查方法专家的任务就是对影响数据质量和成本的不同因素做系统分析。因此，调查方法专家必须从总体调查误差的视角确认哪种搜集数据的方法与调查目的最匹配。

调查方法在不断地转向自访模式吗？有访员帮助的调查在消失吗？

有不少人说访员主导的调查在面临死亡。我们不打算预测人类测量的长远未来。自访模式的好处是有越来越多的可用方法（如电子邮件、网络）。对敏感属性的测量，自访的价值越来越多地得到了证明（参见第5.3.5节）。

尽管自访模式有很多长处，依然有很多情形需要访员才可以成功实施。例如，如果没有抽样框有联系信息（如姓名、地址、电邮），此时，就需要访员去定位样本。如果受访者需要额外的激励才愿意受访，访员的存在就有价值。最后，有些访题，有访员比无访员要更好，特别是需要回忆的开放题。因此，尽管自访模式可以降低敏感议题的测量误差，也可能更加节省成本，在某些情况下，访员对调查执行却起着关键作用。

在问卷调查中，真的会用到不同模式的测量误差吗？

有时候会用到，不过，不那么经常。在第7、8、9章，你学过不同统计模型刻画的是应答形成的过程以及与设计相关的参数估计。用模型估计参数时需要用到设计特征（如交叉访员配置、重访，以及多指标），因此，研究人员需要做额外的工作。如果实施这些特征要求，就会增加调查成本。基于这两点，通常也没有这么做。调查专家在寻找机会搜集测量过程的辅助数据，以获得对不同类型测量误差的估计。计算机辅助方法对一般意义上的调查提供了机会是零边际成本的途径。

如果要获得一项可接受的调查结果，最小应答率要多少？

遗憾的是，即使是同一个调查，针对不同的统计量，答案也可能不同。第6章讨论的无应答率仅仅是无应答误差的一个指标。调查统计中的受访者和非受访者不同，则是无应答误差的另一方面。调查统计的这一特征，一般而言是不可测的。当调查参与的因素与调查统计关联时，在统计上，非受访者应该有与受访者不一样的值。当发生这种情况时，就需要很好的应答率才可以获得低的无应答误差。如果调查参与与统计调查无关，根据定义，则在统计上，受访者与非受访者会有相似的值。在这种情况下，调查估计值会相似，而与应答率无关。大多数调查都不会对此进行说明，因为无法说明。

不管其含义的歧义性，通常都会报告应答率，且通常也会作为数据质量的指标。因此，调查统计的可靠性常常与应答率有关。在缺乏无应答与调查估计之间关系的信息时，说较高的应答率就会有较低的

平均风险是可以的，因为对一个调查而言，无应答对所有估计值都有影响。

应答率的下降意味着调查的死亡吗？

第一，正如第6.2节提到的，家户调查的应答率与那些成功的调查比较的确在下降。第二，政府调查如NCVS和NAEP始终都有很高的应答率。第三，其他国家如英国、荷兰在家户调查中从来就有较低的应答率，不过，用调查作为基本信息搜集工具依然很兴盛。第四，随着人口的变化和社会的变化，搜集数据的调查口径也一直在变。随着单人家庭数量的增长，家户中儿童的数量也在下降，越来越多的妇女也从业了，因此，越来越难在家里找到受访者。调查机构在调查中越来越多地依靠电话。对电访而言，要找到那些难以找到的个体，通常要打20～30个电话。

与利用的下降不同，在过去的10～15年里，在美国甚至世界上，用调查方法搜集数据的趋势在增加。尽管有证据表明在过去曾经管用的搜集数据的方法现在似乎不那么管用了，不过，搜集数据的方法与程序也在发展。

需要像调查方法专家讨论的那样复杂来编写调查访题吗？最终，我们还是把问题作为正常生活的一部分。

日常生活对话中的问题与调查访题之间的关键区别是，调查访题是用来产生可以用于统计归纳应答的。例如，也许你想算一算应答中具有某一特征的百分比：民主党人，左翼，支持堕胎，或去年看过医生。要做到这一点，就要：

1）每个人都回答同样的访题；访题措辞的含义对每位应答者都是一样的（这就需要标准化措辞）。

2）应答应该可以用一致的方法归纳、列表；人们的应答必须是可比较的（这就要求给访题以可选定的应答，在开放访题中，为获得充分信息，指定追问规则）。

从调查方法中发现，访题中微小的措辞变动会在应答中带来很大差别，这一点已经不再是有争议的议题了。从研究中获得的教训尽管不是所有的都已为大众所知，许多也已应用到日常讲话中。需要注意的是，在日常对话中发生误解的情形比在调查中的要少。日常对话中如果出现曲解可以在随后交谈中让双方确认是否误解以及更正。双方共享的基础（通常是之前的多次交流）让他们可以自由地相互澄清。调查访问则更聚焦于互动，需要让受访者在第一次听到访题时有最少的歧义和最多的理解，进而不需要受访者经常与访员澄清。

自访模式对访题的措辞有更多要求。在自访模式中，没有任何机会可以澄清误解。访题的措辞和呈现需要完整地体现研究者的意图。

简而言之，尽管人们总在问问题，不过，他们的遣词造句却不是用于应答列表的通用方式。

多选项测量比单选项测量更好吗？

如果单选项是不够的，则多选项是有吸引力的。在心理测量中，正如我们在第7章讨论的，有时候需要许多选项来建构同一个概念（例如，要用多道不同的计算问题来测量数学能力一样）。这里，研究者们相信建构的焦点是采用单个选项测量是不稳定的。如此，采用多个

选项，就降低了不稳定性。对多选项也有不同的观点，即把多选项看作相互独立的部分组成（例如，收入中包含薪酬、小费、奖金、利息以及股票分红）。因此，多选项测量的结果通常要累计报告。

为什么要对之前调查使用过的访题进行认知测量？

对访题进行认知评估在调查中是非常新的调查研究内容。1990年之前调查用的访题大多都没有进行过认知测试。此外，即使是1990年代开发的访题也不是都进行过认知测试。这样，有许多访题在不同的领域有着长久的应用史，却不符合现在的认知测试标准。

研究者可以探讨之前已经使用过的访题的价值。有时候，替换之前放弃的项目也是有价值的，因为可以测量随时间发生的变化。换个角度看，之前使用过的访题可能说明其经过分析检验具有某种程度的效度。如果缺乏科学评估依据，也可以说，访题曾被用过且有效，进而由此给访题以并非基于标准的信任。

不管采用哪种标准来判断对访题的使用，如果没有经过认知测试，那么在运用到于调查之前，最好还是进行认知测试。

如果之前用过的访题是经过认知测试的，就可能有三个产出。第一，从认知的视角看，经过测试可能是好访题。在这种情况下，每一个人都会确认。第二，从认知的视角看，也许有很大的问题。一旦识别问题，研究者就需要考虑运用访题的理由，例如测量趋势或过去的心理绩效，而不对有严重缺陷的访题进行评价。总之，下一次调查就是未来的趋势测量的基线。在调查中，每次尽量运用最好的访题。第三，如果研究决定要运用在认知测试中发现有严重缺陷的访题，至少

研究和其他使用者要知道访题的局限性。因此，运用这些访题获得的结论可能具有暂时性。

在调查之前测试访题的合理方式是什么？

通常会采用多阶段测试。对研究者而言，如果调查覆盖的是新区域或要研究的新目标总体，对调查程序而言，在目标总体中做几次专题讨论总是值得的。非常重要的第一步是了解目标总体在调查议题上的语汇。

一旦把访题草稿设计好了，从内容到方法的专家讨论也是非常值得的。内容专家可以确认访题会获得需要用于分析的信息，方法专家则可以标出可能有问题的访题。在某些情况下，好访题的设计原则是，应该允许直接修订访题。专家评估标出来的问题也可以再拿到实践中去评估。

在新调查工具开发中，认知测试变成了标准程序。即使是少量的认知访问，少于10，也可能识别出访题理解或应答形成的重要问题。基于测试的修订对调查数据的价值会作出显著的贡献。

此外，实地试调查依然有意义。此外，试调查还有两点可以显著提高结果的价值。第一，对访问调查而言，在试调查中对访谈录音和对访员以及受访者行为进行编码，是获取访问中如何访问以及如何应答的系统数据的低成本方法，也给暴露问题和修订访题提供了访题的量化指标。更进一步说，在试调查中，如果对访题措辞进行随机实验，则可以提供访题措辞效果的经验性数据。对修订之前已经使用过的访题或正在选择访题的某个版本而言，这样的测试更有价值。为了

提供有意义的数据，试调查的对象数量通常要大一些，或许为100～200。不过，由此获得的增量信息说明，这么做是值得的。

认知测试或调查后效度或访题相关偏差测试，哪个更重要？

一方面，对测试某些事而言，我们的工作做得怎么样，最后的测试就是依据某些标准对应答形成的共识。因此，如果在两道访题之间选择，如果人们依据好的建构指标测量达成了共识（参见第8.9节），那么依据科学标准就一定会选择更有效的测量，而不管认知测试的结果如何。如果在应答者层面获得应答偏差指标显示为可忽略误差，就有足够的证据说明是好的测量。

另一方面，更经常的情况是，研究者缺乏好测量的标准来判断其建构如何。在缺乏好的效度标准时，通过评估受访者对访题的理解如何以及是否有能力应答，事实上，这是更直接的降低测量误差的方法。在说明这些强理论之后，我们还得说，从认知的观点来看，并没有足够的证据说明改善访题会让测量更加有效。

因此，坦白地说，回答这个问题，就评估访题而言，认知标准和心理标准都重要。理论上，两者之间密切相关。我们真的需要更多的、更好的研究来帮助我们理解其中的关系。

说访题检验通过了是什么意思？

运用建构效度作为质量标准引出了许多问题（参见第8.9.1节）。有时候，访题的初创者执行了有效的研究（运用外部测量，或多指标），然后经过改良，再次进行效度检验，并报告说检验通过。

第一，通过观察应答与访题之间的关系以及建构访题设计的外部测量来评估访题的效度。的确，很少有所谓“金标准”。此外，与金标准之间的相关系数为1.0的情形，从来也没有发生过。因此，效度就是程度问题。研究者搜集证据引导他们思考至少应答的某些方差是真的来自目标建构。

第二，效度研究的结果（就是效度）还有赖于总体测量。效度研究常常与后来运用访题的总体不一致。

第三，调查方法的研究发现，访题的绩效有时候有赖于访题的语境和搜集数据的方法。

因此，当研究者说他们运用经过效度检验的测量时，就意味着他们在某个时点的某个人群中搜集了一些证据，其应答中的变异性反映了目标建构。与下一个使用者用我们讨论的任意测量提出批评无关。

访员真的是调查误差的重要来源吗？如果是，为什么研究者很少报告这类误差？访员针对调查估计的质量有那么大的影响吗？

显然，对调查访题而言，谁提问，差别应该不大。不过，在研究中如果评估与访员有关的应答，平均组内相关系数会有0.01（Groves，1989；Fowler and Mangione，1990）。看起来，这是一个很小的数。不过，其对标准误的估计可以是：

$$\sqrt{1+\rho_{\text{int}}+(m-1)}$$

式中，m 为一个访员的平均访问数量（参见第9.3节 ）。这意味着，如果访员在每个项目的平均访问数量为50，则一个因素的平均访题的

样本均值标准误。如果平均的 ρ_{int} 为0.01，应该会波动1.22。如果一个访员在项目中平均访问数量为100，就像大型的全国性调查而言，则一个因素的平均访题的样本均值标准误会波动1.41。

研究表明，访员相关的误差可以通过恰当的培训和督导来降低，或者通过设计无需大量访员干预来获得应答的访题而降低。不过，理论上，所有有访员的调查，包括访题，都会受到访员的显著影响，因此研究者通常不能预估访员对统计量的效应。

我的计量经济学家朋友说，在回归模型中，样本权重完全不靠谱。是真的吗？

正如我们在上面的问题中已经说明的，调查统计量是用统计估计来反映调查设计。计量经济学通常采用的是不同的推论框架，即刻画一个无穷过程来产生应变量的值。例如，计量经济学家运用简单回归：

$$y_i = \beta_0 + \beta_i x_i + \varepsilon_i$$

检验x 与y 之间的因果理论。如果模型正确，传统的OLS估计可以产生无偏的 β_i 估计。调查方法专家也许会建议采用选择权重，也就是说，在调查中，年老的人群比年轻的人群具有较高的备选概率。计量经济学家或许认为，如果年龄影响到参数的估计，模型一定有偏，且年龄一定要作为自变量。

调查方法研究专家运用回归模型来归纳框总体中的关系，而不是做因果推论。因此，两者目的不同，路径也不相同。一个实践的路径

是，Dumouchel和Duncan（1983）认为，加权估计与未加权相关系数估计之间区别在于促进对模型的具体化的重新思考。在估计模型参数时，通过权重反映抽样设计和反映整群抽样的标准误之间争论，可以参见Brewer和Mellor（1973）以及Groves（1989，pp. 279-290）。

对一般总体调查数据进行加权以调整备选概率差异在什么时候重要，什么时候不重要？

第4.5.2节和第10.5节讨论过，如果样本成员的备选概率不同，则建议使用选择权重。对覆盖无应答值的调整，取决于对过程的假设，典型的情况是看应答者和非应答者（覆盖到的和没有覆盖到的）在权重组中的调查统计量是不是相等。一般情况下，这个假设用调查数据自身是不可测的。因此，就无法对假设进行评估。不过，可以采用一个实践步骤：看看关键统计量在加权和没加权条件下的标准误。如果结论是两者差不多，就意味着分析获得了较小的估计标准误。

最后，重要的是要注意到统计量对权重的敏感性（例如卡方统计量和标准误估计）会产生误导，如果运用相同的统计软件而没有换过统计软件（参见第4章 和10章 ）。

本书引用的所有抽样和研究都是美国的。这些问题在世界上其他国家也会出现吗？

不。对调查研究者而言，不同的国家可用资源也不相同，对调查测量的反应，在文化上也不相同。在有人口登记的国家，或许有抽样框，也就用不着我们讨论过的复杂抽样设计了。如果一个国家的电话和互联网覆盖了少部分人口，则运用这类基础设施搜集数据就会产生

较大的无观察误差。在识字率较低或邮政系统不大可靠的国家，（邮寄）自访问卷或许无效。如果一个国家的文化不允许与陌生人就利益相关的议题进行诚实的讨论，调查就很难。

我们希望并相信，本书讨论的许多设计原则可以适用于其他环境。不过，对原则的运用要依据情境的可行性与约束进行谨慎的修改。

我想做一个调查。为了做好，我有必要接受调查方法专家那样的训练吗？

正如在第1章讨论过的，调查方法明示的是隐含在不同调查设计中的原则。这些原则通常综合了统计学、心理学、社会学、计算机科学的概念框架。在调查方法领域，如果没有广泛而深入的阅读，就很难掌握如此广博的知识。然而，这些原则并不是说对每一个调查都要选择同样的设计特征。对不同原则的综合运用就要求明确调查的目的。

除了有综合知识以外，为获得最优的综合选择，还要有实地调查知识。因此，在某种意义上，调查方法专家具有这样的理解，也是有价值的。与所有领域一样，新手如果希望立即进入行动，就需要咨询调查专业人士，特别是在调查设计阶段，让关键的决策能从既往的调查研究结论中获益。

参考文献

Abowd, J., and Woodcock, S. (2001), “Disclosure Limitation in Longitudinal Linked Data.” In Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L., eds., *Confidentiality*, *Disclosure*, *and Data Access*: *Theory and Practical Applications for Statistical Agencies*, Amsterdam, North-Holland, pp. 215-277.

Abowd, J. M., Stinson, M., and Benedetto, G., “Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project,” November 2006, http://www.census.gov/sipp/synth_data.html.

Alreck, P., and Settle, R. (1995), *The Survey Research Handbook*, New York: McGraw-Hill.

Alwin, D. (2007), *Margins of Error*, New York: Wiley.

American Association for Public Opinion Research (2000), *Standard Definitions*: *Final Dispositions of Case Codes and Outcome Rates for Surveys*, Ann Arbor, Michigan: AAPOR.

American Psychological Association (2003), “Psychological Research Online: Opportunities and

Challenges," Working Paper Version 3/31/03, Washington, DC: American Psychological Association.

Anderson, M. (1990), *The American Census : A Social History* , New Haven: Yale University Press.

Andrews, F. (1984), "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach," *Public Opinion Quarterly* , 48, pp. 409-422.

Aneshensel, C., Frerichs, R., Clark, V., and Yokopenic, P. (1982), "Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews," *Public Opinion Quarterly* , 46, pp. 110-121.

Aquilino, W. (1992), "Telephone Versus Face-to-Face Interviewing for Household Drug Use Surveys," *International Journal of the Addictions* , 27, pp. 71-91.

Aronson, E., and Carlsmith, J. (1969), "Experimentation in Social Psychology," in Lindzey, G., and Aronson, E. (eds.), *Handbook of Social Psychology* , 2nd ed., vol. 2, pp. 1-79, Reading, MA: Addison-Wesley.

Asch, S. (1956), "Studies of Independence and Conformity," *Psychological Monographs* , 70, No. 416.

Atrostic, B., and Burt, G. (1999), "What Have We Learned and a Framework for the Future," in *Seminar on Interagency*

Coordination and Cooperation , Statistical Policy Working Paper 28, Washington, DC: Federal Committee on Statistical Methodology.

Babbie, E. (1990), *Survey Research Methods* (2nd edition), Belmont, CA: Wadsworth.

Babbie, E. (2004), *The Practice of Social Research* , Belmont, CA: Wadsworth.

Bahrick, H., Bahrick, P., and Wittlinger, R. (1975), “Fifty Years of Memory for Names and Faces: A Cross Sectional Approach,” *Journal of Experimental Psychology: General* , 104, pp. 54-75.

Barsalou, L. (1988), “The Content and Organization of Autobiographical Memories,” in Neisser, U., and Winograd, E. (eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory* , pp. 193-243, Cambridge, U. K.: Cambridge University Press.

Bates, N., Dahlhammer, J., and Singer, E. (2008), “Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Nonresponse,” *Journal of Official Statistics* , 24, pp. 591-612.

Battaglia, M., Link, M., Frankel, M., Osborn, L., and Mokdad; A. (2008), “An Evaluation of Respondent Selection

Methods for Household Mail Surveys," *Public Opinion Quarterly* , 72, pp. 459-469.

Beatty, P. (2004), "The Dynamics of Cognitive Interviewing," in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Beatty, P., and Herrmann, D. (2002), "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse," in Groves, R., Dillman, D., Eltinge J., and Little, R. (eds.), *Survey Nonresponse* , pp. 71-85, New York: Wiley.

Beebe, T., Harrison, P., McRae, J., Anderson, R., and Fulkerson, J. (1998), "An Evaluation of Computer-Assisted Self-Interviews in a School Setting," *Public Opinion Quarterly* , 62, pp. 623-632.

Béland, Y., and St-Pierre, M. (2008), "Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI," in Lepkowski, J., Tucker, C., Brick, J., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (eds.), *Advances in Telephone Survey Methodology* , New York: Wiley, pp. 297-314.

Belli, R. (1998), "The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements

in the Quality of Retrospective Reports in Surveys," *Memory* , 6, pp. 383-406.

Belli, R., Shay, W., and Stafford, F. (2001), "Event History Calendars and Question Lists," *Public Opinion Quarterly* , 65, pp. 45-74.

Belli, R., Schwarz, N., Singer, E., and Talarico, J. (2000), "Decomposition Can Harm the Accuracy of Behavioral Frequency Reports," *Applied Cognitive Psychology* , 14, pp. 295-308.

Belson, W. (1981), *The Design and Understanding of Survey Questions* , Aldershot: Gower Publishing.

Belson, W. (1986), *Validity in Survey Research* , Aldershot: Gower Publishing.

Bern, D., and McConnell, H. (1974), "Testing the Self-Perception Explanation of Dissonance Phenomena: On the Salience of Premanipulation Attitudes," *Journal of Personality and Social Psychology* , 14, pp. 23-31.

Berk, R. (1983), "An Introduction to Sample Selection Bias in Sociological Data," *American Sociological Review* , 48, pp. 386-398.

Berlin, M., Mohadjer, L., Waksberg, 1., Kolstad, A., Kirsch, I., Rock, D., and Yamamoto, K. (1992), "An

Experiment in Monetary Incentives," in *Proceedings of the Survey Research Methods Section of the American Statistical Association* , pp. 393-398, Washington, DC: American Statistical Association.

Berman, J., McCombs, H., and Boruch, R. (1977), "Notes on the Contamination Method: Two Small Experiments in Assuring Confidentiality of Response," *Sociological Methods and Research* , 6, pp. 45-63.

Bernard, C. (1989), *Survey Data Collection Using Laptop Computers* , Paris: Institut National de la Statistique et des Études Economiques (INSEE), Report No. 01/C520.

Berscheid, E., Baron, R., Dermer, M., and Libman, M. (1973), "Anticipating Informed Consent: An Empirical Approach," *American Psychologist* , 28, pp. 913-925.

Bethlehem, J. (1998), "The Future of Data Editing," in Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls II, W., and O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection* , pp. 201-222, New York: Wiley.

Bethlehem, J. (2002), "Weighting Nonresponse Adjustments Based on Auxiliary Information," in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 275-288, New York: Wiley.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality* , New York: Wiley.

Biemer, P. and Stokes, L. (1991), “Approaches to the Modeling of Measurement Errors,” in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 487-516, New York: Wiley.

Biemer, P., Herget, D., Morton, J., and Willis, G. (2003), “The Feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI),” in *Proceedings of the Survey Research Methods Section of the American Statistical Association* , pp. 1068-1073, Washington, DC: American Statistical Association.

Billiet, J., and Loosveldt, G. (1988), “Interviewer Training and Quality of Responses,” *Public Opinion Quarterly* , 52, pp. 190-211.

Bishop, G., Hippler, H., Schwarz, N., and Strack, F. (1988), “A Comparison of Response Effects in Self-Administered and Telephone Surveys,” in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 321-340, New York: Wiley.

Bishop, G., Oldendick, R., and Tuchfarber, A. (1986), “Opinions on Fictitious Issues: The Pressure to Answer

Survey Questions," *Public Opinion Quarterly* , 50, pp. 240-250.

Blair, E., and Burton, S. (1987), "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions," *Journal of Consumer Research* , 14, pp. 280-288.

Blumberg, S., Luke, J., Cynamon, M., and Frankel, M. (2008), "Recent Trends in Household Telephone Coverage in the United States." In Lepkowski, J., Tucker, C, Brick, J., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (eds.), *Advances in Telephone Survey Methodology* , pp. 56-86, New York: Wiley.

Blumberg, S., and Luke, J. (2008), *Wireless Substitution: Early Release of Estimates From the National Health Interview Survey* , *July-December 2007* , www.cdc.gov/nchs.

Bogen, K. (1996), "The Effect of Questionnaire Length on Response Rates: A Review of the Literature," in *Proceedings of the Survey Research Methods Section of the American Statistical Association* , Alexandria, VA: American Statistical Association, pp. 1020-1025.

Booth, C (1902—1903), *Life and Labour of the People of London* , London and New York: MacMillan Co.

Bosnjak, M., and Tuten, T. (2001), "Classifying Response Behaviors in WebBased Surveys," *Journal of Computer-Mediated*

Communication , 6(3) (http://jcmc.indiana.edu).

Botman, S., and Thornberry, O. (1992), "Survey Design Features Correlates of Nonresponse," in *Proceedings of the Survey Research Methods Section of the American Statistical Association* , pp. 309-314, Alexandria, VA: American Statistical Association.

Boyle, J., Kilpatrick, D., Acinerno, R., Ruggiero, K., Resnick, H., Galea, S., Koenan, K., and Galernter, J. (2007), "Biological Specimen Collection in an RDD Telephone Survey: 2004 Florida Hurricanes Gene and Environment Study," in *Proceedings of the 9th Conference on Health Survey Research Methods* , Hyattsville, MD: National Center for Health Statistics.

Bradburn, N., Sudman, S., and Associates (1979), *Improving Interview Method and Questionnaire Design* , San Francisco: Jossey-Bass.

Brehm, J. (1993), *The Phantom Respondents* : *Opinion Surveys and Political Representation* , Ann Arbor: University of Michigan Press.

Brewer, K., and Mellor, R. (1973), "The Effect of Sample Structure on Analytical Surveys," *Australian Journal of Statistics* , 15, pp. 145-152.

Brick, M., Montaquila, J., and Scheuren, F. (2002), "Estimating Residency Rates," *Public Opinion Quarterly* , 66, pp. 18-39.

Brϕgger, J., Bakke, P., Eide, G., and Guldvik, A. (2002), "Comparison of Telephone and Postal Survey Modes on Respiratory Symptoms and Risk Factors." *American Journal of Epidemiology* , 155, pp. 572-576.

Brunner, G., and Carroll, S. (1969), "The Effect of Prior Notification on the Refusal Rate in Fixed Address Surveys," *Journal of Marketing Research* , 9, pp. 42-44.

Bureau of the Census (2002), *Voting and Registration in the Election of November* 2000, *Current Population Reports* , P20-542, Washington, DC: U.S. Bureau of the Census.

Burton, S., and Blair, E. (1991), "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys," *Public Opinion Quarterly* , 55, pp. 50-79.

Campanelli, P., Thomson, K., Moon, K., and Staples, T. (1997), "The Quality of Occupational Coding in the UK," in Lyberg L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality* , pp. 437-457, New York: Wiley.

Cannell, C., and Fowler, F. (1964), “A Note on Interviewer Effect in SelfEnumerative Procedures,” *American Sociological Review* , 29, p. 276.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., and Miller, P. (1987), “An Experimental Comparison of Telephone and Personal Health Interview Surveys,” *Vital and Health Statistics* , Series 2, 106, Washington, DC: Government Printing Office.

Cannell, C., Marquis, K., and Laurent, A. (1977), “A Summary of Studies,” *Vital and Health Statistics* , Series 2, 69, Washington, DC: Government Printing Office.

Cannell, C., Miller, P., and Oksenberg, L. (1981), “Research on Interviewing Techniques,” in Leinhardt, S. (ed.), *Sociological Methodology* 1981, pp. 389-437, San Francisco: Jossey-Bass.

Cantor, D., and Esposito, J. (1992), “Evaluating Interviewer Style for Collecting Industry and Occupation Information,” *Proceedings of the Section on Survey Research Methods, American Statistical Association* , pp. 661-666.

Catlin, O., and Ingram, S. (1988), “The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing,” in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J.

(eds.), *Telephone Survey Methodology* , pp. 437-450, New York: Wiley.

Centers for Disease Control (2005), *BRFSS User’s Guide* , http://www.cdc.gov/brfss/pdf/userguide.pdf.

Centers for Disease Control (2008), *BRFSS Questionnaires* , http://www.cdc.gov/brfss/questionnaires/pdf-ques/2008brfss.pdf.

Citro, C., Ilgen, D., and Marrett, C. (2003), *Protecting Participants and Facilitating Social and Behavioral Sciences Research* , Washington, DC: National Academy Press.

Cochran, W. (1961), “Comparison of Methods for Determining Stratum Boundaries,” *Bulletin of the International Statistical Institute* , 38, pp. 345-358.

Cochran, W. (1977), *Sampling Techniques* , New York: Wiley.

Cohany, S., Polivka, A., and Rothgeb, J. (1994), “Revisions in the Current Population Survey Effective January 1994,” *Employment and Earnings* , February, pp. 13-37.

Colledge, M., and Boyko, E. (2000), *UN/ECE Work Session on Statistical Metadata (METIS)* , Washington, November 28-30.

Collins, C. (1975), "Comparison of Month-to-Month Changes in Industry and Occupation Codes with Respondent's Report of Change: CPS Mobility Study." Response Research Staff Report 75-76, May 15, 1975, U.S. Bureau of the Census.

Collins, M., and Courtenay, G. (1985), "A Comparison Study of Field and Office Coding," *Journal of Official Statistics* , 1, pp. 221-227.

Conrad, F., and Blair, J. (1996), "From Impressions to Data: Increasing the Objectivity of Cognitive Interviews," in *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association* , Alexandria, VA: American Statistical Association, pp. 1-10.

Conrad, F, and Schober, M. (2000), "Clarifying Question Meaning in a Household Telephone Survey," *Public Opinion Quarterly* , 64, pp. 1-28.

Conrad, F., and Schober, M. (2008), *Envisioning the Survey Interview of the Future* , New York: Wiley.

Conrad, F., Brown, N., and Cashman, E. (1998), "Strategies for Estimating Behavioral Frequency in Survey Interviews," *Memory* , 6, pp. 339-366.

Converse, J. (1987), *Survey Research in the United States* , Berkeley: University of California Press.

Converse, J., and Presser, S. (1986), *Survey Questions : Handcrafting the Standardized Questionnaire* , Thousand Oaks, CA: Sage.

Conway, M. (1996), “Autobiographical Knowledge and Autobiographical Memories,” in Rubin, D. (ed.), *Remembering Our Past* , pp. 67-93, Cambridge, U. K.: Cambridge University Press.

Cook, C., Heath, F, and Thompson, R. (2000), “A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys,” *Educational and Psychological Measurement* , 60, pp. 821-836.

Couper, M. (1996), “Changes in Interview Setting under CAPI,” *Journal of Official Statistics* , 12, pp. 301-316.

Couper, M. (2000), “Web Surveys: A Review of Issues and Approaches,” *Public Opinion Quarterly* , 64, pp. 464-494.

Couper, M. (2001), “The Promises and Perils of Web Surveys,” in Westlake, A. et al. (eds.), *The Challenge of the Internet* , pp. 35-56, London: Association for Survey Computing.

Couper, M. (2008a), “Technology and the Survey Interview/Questionnaire,” in Schober, M. F, and Conrad, F. G. (eds.), *Envisioning the Survey Interview of the Future* , New York: Wiley, pp. 58-76.

Couper, M. (2008b), *Designing Effective Web Surveys.* New York: Cambridge University Press.

Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls, w., and O'Reilly, J. (1998), *Computer Assisted Survey Information Collection* , New York: Wiley.

Couper, M., Blair, J., and Triplett, T. (1999), "A Comparison of Mail and E-Mail For a Survey of Employees in Federal Statistical Agencies," *Journal of Official Statistics* , 15, pp. 39-56.

Couper, M., and Groves, R. (2002), "Introductory Interactions in Telephone Surveys and Nonresponse," in Maynard, D., Houtkoop-Steenstra, H., Schaeffer, N., and van der Zouwen, J. (eds.), *Standardization and Tacit Knowledge* : *Interaction and Practice in the Survey Interview* , pp. 161-177, New York: Wiley.

Couper, M., Hansen, S., and Sadosky, S. (1997), "Evaluating Interviewer Use of CAPI Technology, in Lyberg, L., Biemer, P., Collins, M., Dippo, C., and Schwarz, N. (eds.), *Survey Measurement and Process Quality* , pp. 267-286, New York: Wiley.

Couper, M., Kapteyn, A., Schonlau, M., and Winter, J. (2007), "Noncoverage and Nomesponse in an Internet Survey," *Social Science Research* , 36, pp. 131-148.

Couper, M., and Nicholls II, W. (1998), “The History and Development of Computer Assisted Survey Information Collection Methods,” in Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls II, W., and O’Reilly, J. (eds.), *Computer Assisted Survey Information Collection* , pp. 1-22, New York: Wiley.

Couper, M., and Rowe, B. (1996), “Computer-Assisted Self-Interviews,” *Public Opinion Quarterly* , 60, pp. 89-105.

Couper, M., Singer, E., Comad, F. G., and Groves, R. (2008), “Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation.” *Journal of Official Statistics* , 24, pp. 255-75.

Couper, M., Singer, E., and Kulka, R. (1998), “Participation in the 1998 Decennial Census: Politics, Privacy, Pressures,” *American Politics Quarterly* , 26, pp. 59-80.

Cronbach, L. (1951), “Coefficient Alpha and the Internal Structure of Tests,” *Psychiatrika* , 16, pp. 297-334.

Cronbach, L., and Meehl, P. (1955), “Construct Validity in Psychological Tests,” *Psychological Bulletin* , 52, pp. 281-302.

Csikszentmihalyi, M., and Csikszentmihalyi, I. (eds.) (1988), *Optimal Experience : Psychological Studies in Flow of Consciousness* , New York: Cambridge University Press.

Curtin, R. (2003), "Unemployment Expectations: The Impact of Private Information on Income Uncertainty," *Review of Income and Wealth* , 49, pp. 539-554.

Curtin, R. (2003), *Surveys of Consumers : Sample Design* , http://www.sca.isr.umich.edu/.

Curtin, Richard T. (2003), *Surveys of Consumers : Survey Description* , http://www.sca.isr.umich.edu/.

de la Puente, M. (1993), "A Multivariate Analysis of the Census Omission of Hispanics and Non-Hispanic Whites, Blacks, Asians and American Indians: Evidence from Small Area Ethnographic Studies," in *Proceedings of the Survey Research Methods Section , American Statistical Association* , pp. 641-646, Alexandria, VA: American Statistical Association.

de Leeuw, E. (1992), *Data Quality in Mail , Telephone and Face-to-Face Surveys* , Amsterdam: TT-Publikaties.

de Leeuw, E. (2005), "To Mix or Not to Mix Data Collection Modes in Surveys," *Journal of Official Statistics* , 21, pp. 233-255.

de Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., and Lensvelt-Mulders, G. (2007), “The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis,” *Public Opinion Quarterly* , 71, pp. 413-443.

de Leeuw, E., and de Heer, W. (2002), “Trends in Household Survey Nonresponse: A Longitudinal and International Comparison,” Chapter 3 in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 41-54, New York: Wiley.

de Leeuw, E., and van der Zouwen, J. (1988), “Data Quality in Telephone and Face-to-Face Surveys: A Comparative Meta-Analysis,” in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 283-299, New York: Wiley.

DeMaio, T., and Landreth, A. (2004), “Do Different Cognitive Interview Methods Produce Different Results?” in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Deming, W. (1950), *Some Theory of Sampling* , New York: Dover.

Denscombe, M. (2008), “The Length of Responses to Open-Ended Questions; A Comparison of Online and Paper

Questionnaires in Terms of a Mode Effect," *Social Science Computer Review* , 26, pp. 359-368.

Deutskens, E., de Ruyter, K., and Wetzels, M. (2006), "An Assessment of Equivalence between Online and Mail Surveys in Service Research," *Journal of Service Research* , 8, pp. 346-355.

Dielman, L., and Couper, M. (1995), "Data Quality in CAPI Surveys: Keying Errors," *Journal of Official Statistics* , 11, pp. 141-146.

Dillman, D. (1978), *Mail and Telephone Surveys* : *The Total Design Method* , New York: Wiley.

Dillman, D., Eltinge, J., Groves, R., and Little, R. (2002), "Survey Nonresponse in Design, Data Collection and Analysis," in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 3-26, New York: Wiley.

Dillman, D., Sinclair, M., and Clark, J. (1993), "Effects of Questionnaire Length, Respondent Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys," *Public Opinion Quarterly* , 57, pp. 289-304.

Dillman, D., Smyth, J., and Christian, L. (2009), *Internet, Mail* , *and Mixed-Mode Surveys* : *The Tailored Design*

Method. New York: Wiley.

Dillman, D., and Tarnai, J. (1988), "Administrative Issues in Mixed-Mode Surveys," in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 509-528, New York: Wiley.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001), *Confidentiality* , *Disclosure* , *and Data Access* : *Theory and Practical Applications for Statistical Agencies* , Amsterdam: North-Holland.

DuMouchel, W, and Duncan, G. (1983), "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," *Journal of the American Statistical Association* , 78, pp. 535-543.

Economic Classification Policy Committee (1993), *Issues Paper No.* 1: *Conceptual Issues* , Washington, DC: Bureau of Economic Analysis.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., and Kwan, I. (2002), "Increasing Response Rates to Postal Questionnaires: Systematic Review," *British Medical Journal* , 324, pp. 1183-1192.

Edwards, W., Winn, D., Kurlantzick, V., Sheridan, S., Berk, M., Retchin, S., and Collins, J. (1994), "Evaluation

of National Health Interview Survey Diagnostic Reporting," *Vital and Health Statistics*, Series 2, No. 120, Hyattsville, MD: National Center for Health Statistics.

Edwards, W., and Cantor, D. (1991), "Toward a Response Model in Establishment Surveys," in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys*, pp. 221-236, New York: Wiley.

Edwards, W, Winn, D., and Collins, J. (1996), "Evaluation of 2-week Doctor Visit Reporting in the National Health Interview Survey," in *Vital and Health Statistics*, Series 2, No. 122, Hyattsville, MD: National Center for Health Statistics.

Ekholm, A, and Laaksonen, S. (1991), "Weighting via Response Modeling in the Finnish Household Budget Survey," *Journal of Official Statistics*, 7, pp. 325-377.

Ericsson, K., and Simon, H. (1980), "Verbal Reports as Data," *Psychological Review*, 87, pp. 215-251.

Ericsson, K, and Simon, H. (1984), *Protocol Analysis: Verbal Reports as Data*, Cambridge, MA: MIT Press.

Erlich, J., and Riesman, D. (1961), "Age and Authority in the Interview," *Public Opinion Quarterly*, 24, pp. 99-114.

Etter, J. F., Perneger, T., and Ronchi, A (1998), "Collecting Saliva Samples by Mail," *American Journal of Epidemiology* , 147, pp. 141-146.

Faden, R., and Beauchamp, T. (1986), *A History and Theory of Informed Consent* , New York: Oxford University Press.

Federal Committee on Statistical Methodology (1994), *Working Paper* 22: *Report on Statistical Disclosure Limitation Methods* , Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC.

Federal Register (1991), "Federal Policy for the Protection of Human Subjects," June 18, pp. 280002-280031.

Fellegi, I., and Holt, T. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association* , 71, pp. 17-35.

Fellegi, I. (1964), "Response Variance and Its Estimation," *Journal of the American Statistical. Association* , 59, pp. 1016-1041.

Felsö, F., Theeuwes, J., and Wagner, G. (2001), "Disclosure Limitations Methods in Use: Results of a Survey," in Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (eds.), *Confidentiality* , *Disclosure* , *and Data Access* :

Theory and Practical Applications for Statistical Agencies , Amsterdam: North-Holland/Elsevier.

Fienberg, S. (1994), “A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality,” Carnegie Mellon University, Department of Statistics, Technical Report 611, Pittsburgh, PA: Carnegie Mellon University.

Fienberg, S., and Makov, U. (1998), “Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data,” *Journal of Official Statistics* , 14, pp. 385-397.

Fienberg, S., Steele, R., and Makov, U (1996), “Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Log-Linear Models,” in *Proceedings of the Bureau of the Census* 1996 *Annual Research Conference* , pp. 87-105, Washington, DC: U.S. Bureau of the Census.

Forsman, G., and Schreiner, I. (1991), “The Design and Analysis of Reinterview: An Overview,” in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 279-301, New York: Wiley.

Forsyth, B., and Lessler, J. (1992), “Cognitive Laboratory Methods: A Taxonomy,” in Biemer, P., Groves, R.,

Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 393-418, New York: Wiley.

Forsyth, B., Rothgeb, J., and Willis, G. (2004), “Does Pretesting Make a Difference?” in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Fowler, F. (1992), “How Unclear Terms Affect Survey Data,” *Public Opinion Quarterly* , 56, pp. 218-231.

Fowler, F. (1995), *Improving Survey Questions* , Thousand Oaks, CA: Sage Publications.

Fowler, F. (2001), *Survey Research Methods* , (3rd Edition), Thousand Oaks, CA: Sage Publications.

Fowler, F. (2004), “Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies,” in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Fowler, F., and Cannell, C. (1996), “Using Behavioral Coding to Identify Cognitive Problems with Survey Questions,” in Schwarz, N., and Sudman, S. (eds.), *Answering Questions* , pp. 15-36, San Francisco: Jossey-Bass.

Fowler, F., Gallagher, P., Stringfellow, V., Zaslavsky, A., Thompson, J., and Cleary, P. (2002), “Using Telephone

Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members," *Medical Care* , 40, pp. 190-200.

Fowler, F., and Mangione, T. (1990), *Standardized Survey Interviewing* : *Minimizing Interviewer-Related Error* , Beverly Hills, CA: Sage Publications.

Frankel, L. (1983), "The Report of the CASRO Task Force on Response Rates," in Wiseman, F. (ed.), *Improving Data Quality in a Sample Survey* , pp. 1-11, Cambridge, MA: Marketing Science Institute.

Frey, J. (1986), "An Experiment with a Confidentiality Reminder in a Telephone Survey," *Public Opinion Quarterly* , 50, pp. 267-269.

Fuchs, M., Couper, M., and Hansen, S. (2000), "Technology Effects: Do CAPI Interviews Take Longer?" *Journal of Official Statistics* , 16, pp. 273-286.

Gardner, G. (1978), "Effects of Federal Human Subjects Regulations on Data Obtained in Environmental Stress Research," *Journal of Personality and Social Psychology* , 36, pp. 628-634.

Gaziano, C. (2005), "Comparative Analysis of Within-Household Respondent Selection Techniques," *Public Opinion Quarterly* , 69, pp. 124-157.

Gfroerer, J., Eyerman, J., and Chromy, J. (eds.) (2002), *Redesigning an Ongoing National Household Survey: Methodological Issues* , DHHS Pub. No. SMA 03-3768, Rockville, MD: SAMHSA.

Gottfredson, M., and Hindelang, M. (1977), “A Consideration of Telescoping and Memory Decay Biases in Victimization Surveys,” *Journal of Criminal Justice* , 5, pp. 205-216.

Goyder, J. (1985), “Face-to-Face Interviews and Mail Questionnaires: The Net Difference in Response Rate,” *Public Opinion Quarterly* , 49, pp. 234-252.

Goyder, J. (1987), *The Silent Minority* : *Nonrespondents on Sample Surveys* , Boulder, CO: Westview Press.

Graesser, A., Bommareddy, S., Swarner, S., and Golding, J. (1996), “Integrating Questionnaire Design with a Cognitive Computational Model of Human Question Answering,” in Schwarz, N., and Sudman, S. (eds.), *Answering Questions* , pp. 143-174, San Francisco: Jossey-Bass.

Graesser, A., Kennedy, T., Wiemer-Hastings, P., and Ottati, V. (1999), “The Use of Computational Cognitive Models to Improve Questions on Surveys and Questionnaires,” in Sirken, M., et al. (eds.), *Cognition in Survey Research* , pp. 199-216, New York: Wiley.

Graham, D. (1984), “Response Errors in the National Crime Survey: July 1974—June 1976”, in Lehnen, R., and Skogan, W. (eds.), *The National Crime Survey : Working Papers* , pp. 58-64, Washington, DC: Bureau of Justice Statistics.

Griffin, D., Fischer, D., and Morgan, M. (2001), “Testing an Internet Response Optiqn for the American Community Survey,” Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Quebec, May.

Groves, R. (1979), “Actors and Questions in Telephone and Personal Interview Surveys,” *Public Opinion Quarterly* , 43, pp. 190-205.

Groves, R. (1989), *Survey Errors and Survey Costs* , New York: Wiley.

Groves, R. (2006), “Nonresponse Rates and Nonresponse Bias in Household Surveys,” *Public Opinion Quarterly* , 70, pp. 646-675.

Groves, R., and Couper, M. (1998), *Nonresponse in Household Interview Surveys* , New York: Wiley.

Groves, R., and Kahn, R. (1979), *Surveys by Telephone : A National Comparison with Personal Interviews* , New York: Academic Press.

Groves, R., and Lyberg, L. (1988), "An Overview of Non-Response Issues in Telephone Surveys," in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 191-212, New York: Wiley.

Groves, R., and Magilavy, L. (1980), "Estimates of Interviewer Variance in Telephone Surveys," in *Proceedings of the Survey Research Methods Section of the American Statistical Association* , pp. 622-627, Alexandria, VA: American Statistical Association.

Groves, R., and Peytcheva, E. (2008) "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly* , 72, pp. 167-189.

Groves, R., Presser, S., and Dipko, S. (2004), "The Role of Topic Salience in Survey Participation Decisions," *Public Opinion Quarterly* , 68, pp. 2-31.

Groves, R., Singer, E., and Coming, A. (2000), "Leverage-Salience Theory of Survey Participation: Description and an Illustration," *Public Opinion Quarterly* , 64, pp. 299-308.

Groves, R., Singer, E., Coming, A., and Bowers, A. (1999), "A Laboratory Approach to Measuring the Joint Effects of Interview Length, Incentives, Differential

Incentives, and Refusal Conversion on Survey Participation," *Journal of Official Statistics*, 15, pp. 251-268.

Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.) (2002), *Survey Nonresponse*, New York: Wiley.

Groves, R., Wissoker, D., Greene, L., McNeeley, M., and Montemarano, D. (2000), "Common Influences on Noncontact Nonresponse Across Household Surveys: Theory and Data," paper presented at the annual meetings of the American Association for Public Opinion Research.

Groves, R., Couper, M., Presser, S., Singer, E., Tourangeau, R., Piani Acosta, G., and Nelson, L. (2006), "Experiments in Producing Nonresponse Bias," *Public Opinion Quarterly*, 70, pp. 720-736.

Guarino, J., Hill, J., and Woltman, H. (2001), *Analysis of the Social Security Number-Notification Component of the Social Security Numbers, Privacy Attitudes, and Notification Experiment*, Washington, DC: U.S Census Bureau.

Gwartney, P. (2007), *The Telephone Interviewer's Handbook: How to Conduct Standardized Conversations*, New York: Wiley.

Haney, C., Banks, C., and Zimbardo, P. (1973), "Interpersonal Dynamics in a Simulated Prison,"

International Journal of Criminology and Penology , 1, pp. 69-97.

Hansen, B., and Hansen, K. (1995), “Academic and Scientific Misconduct: Issues for Nurse Educators,” *Journal of Professional Nursing* , 11, pp. 31-39.

Hansen, M., Hurwitz, W., and Bershad, M. (1961), “Measurement Errors in Censuses and Surveys,” *Bulletin of the International Statistical Institute* , 38, pp. 359-374.

Hansen, M., Hurwitz, W., and Madow, W. (1953), *Sample Surveys Methods and Theory* , Vols. I and II, New York: Wiley.

Hansen, S., and Couper, M. (2004), “Usability Testing as a Means of Evaluating Computer-Assisted Survey Instruments,” in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Harkness, J., Vijver, F., and Mohler, P. (2002), *Cross-Cultural Survey Methods* , New York: Wiley.

Hartley, H. (1962), “Multiple Frame Surveys,” in *Proceedings of the Social Statistics Section* , *American Statistical Association* , pp. 203-206, Alexandria, VA: American Statistical Association.

Hawala, S. (2000), “On the Variation of the Percent of Uniques in a Microdata Sample and the Sample Size.

Statistical Research Division, United States Census Bureau (unpublished memo).

Heberlein, T., and Baumgartner, R. (1978), "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature," *American Sociological Review*, 43, pp. 447-462.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, pp. 153-161.

Henson, R., Roth, A. and Cannell, C. (1977), "Personal Versus Telephone Interviews: The Effects of Telephone Re-Interviews on Reporting of Psychiatric Symptomatology," in Cannell, C., Oksenberg, L., and Converse, J. (eds.), *Experiments in Interviewing Techniques*: *Field Experiments in Health Reporting*, 1971—1977, pp. 205-219, Hyattsville, MD: U.S. Department of Health, Education and Welfare, National Center for Health Services Research.

Hill, C., Donelan, K., and Frankel, M. (1999), "Within-Household Respondent Selection in an RDD Telephone Survey: A Comparison of Two Methods," paper presented at the 1999 meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Hippler, H., Schwarz, N., and Sudman, S. (1987), *Social Information Processing and Survey Methodology*, New York:

Springer-Verlag.

Hochstim, J. (1967), “A Critical Comparison of Three Strategies of Collecting Data from Households,” *Journal of the American Statistical Association* , 62, pp. 976-989.

Holmberg, A., Lorenc, B., and Werner, P. (2008), “Optimal Contact Strategy in a Mail-and-Web Mixed Mode Survey,” Paper presented at the General Online Research Conference (GOR’08), Hamburg, March.

Horn, J. (1978), “Is Informed Deceit the Answer to Informed Consent?” *Psychology Today* , May, pp. 36-37.

Horvitz, D., Weeks, M., Visscher, W., Folsom, R., Massey, R., and Ezzati, T. (1990), “A Report of the Findings of the National Household Seroprevalence Survey Feasibility Study,” *Proceedings of the American Statistical Association* , *Survey Research Methods Section* , pp. 150-159.

Houtkoop-Steenstra, H. (2000), *Interaction and the Standardized Survey Interview* : *The Living Questionnaire* , Cambridge, U.K.: Cambridge University Press.

Hox, J., and de Leeuw, E. (2002), “The Influence of Interviewers’ Attitude and Behavior on Household Survey Nonresponse: An International Comparison,” in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 103-120, New York: Wiley.

Hox, J., and de Leeuw, E. (1994), “A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys: Applying Multilevel Modeling to Meta-Analysis,” *Quality and Quantity* , 28, pp. 329-344.

Hughes, A., Chromy, J., Giacoletti, K., and Odom, D. (2002), “Impact of Interviewer Experience on Respondent Reports of Substance Use,” in Gfroerer, J., Eyerman, J., and Chromy, J. (eds.), *Redesigning an Ongoing National Household Survey* , pp. 161-184, Washington, DC: Substance Abuse and Mental Health Services Administration.

Humphreys. L. (1970), *Tearoom Trade* : *Impersonal Sex in Public Places* , Chicago: Aldine.

Huttenlocher, J., Hedges, L., and Bradburn, N. (1990), “Reports of Elapsed Time: Bounding and Rounding Processes in Estimation,” *Journal of Experimental Psychology* : *Learning* , *Memory* , *and Cognition* , 16, pp. 196-213.

Hyman, H., Cobb, J., Feldman, J., and Stember, C. (1954), *Interviewing in Social Research* , Chicago：University of Chicago Press.

Iannacchione, V, Staab, J., and Redden, D. (2003), “Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey,” *Public Opinion Quarterly* , 67, pp. 202-210.

ISR Survey Research Center (1999), *Center Survey* , April, pp. 1,3.

Inter-university Consortium for Political and Social Research (2001), *National Crime Victimization Survey* , 1992—1999, ICPSR 6406, Ann Arbor: ICPSR.

Jabine, T, King, K., and Petroni, R. (1990), *SIPP Quality Profile* , Washington, DC: U.S. Bureau of the Census.

Jabine, T, Straf, M., Tanur, J., and Tourangeau, R. (eds.) (1984), *Cognitive Aspects of Survey Methodology* : *Building a Bridge between Disciplines* , Washington, DC: National Academy Press.

Jenkins, C., and Dillman, D. (1997), "Towards a Theory of Self-Administered Questionnaire Design," in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality* , pp. 165-196, New York: Wiley.

Jobe, J., and Mingay, D. (1989), "Cognitive Research Improves Questionnaires," *American Journal of Public Health* , 79, pp. 1053-1055.

Johnson, T., Hougland, J., and Clayton, R. (1989), "Obtaining Reports of Sensitive Behaviors: A Comparison of Substance Use Reports from Telephone and Face-to-Face Interviews," *Social Science Quarterly* , 70, pp. 174-183.

Jordan, L., Marcus, A., and Reeder, L. (1980), “Response Styles in Telephone and Household Interviewing: A Field Experiment,” *Public Opinion Quarterly* , 44, pp. 210-222.

Junn, J. (2001), “The Influence of Negative Political Rhetoric: An Experimental Manipulation of Census 2000 Participation,” paper presented at the Midwest Political Science Association, Chicago.

Juster, F., and Smith, J. (1997), “Improving the Quality of Economic Data: Lessons from the HRS and AHEAD,” *Journal of the American Statistical Association* , 92, pp. 1268-1278.

Juster, F., and Suzman, R. (1995), “An Overview of the Health and Retirement Study,” *Journal of Human Resources* , 30, pp. S7-S56.

Kahn, R., and Cannell, C. (1958), *Dynamics of Interviewing* , New York: Wiley.

Kallick-Kaufrnan, M. (1979), “The Micro and Macro Dimensions of Gambling in the United States,” *The Journal of Social Issues* , 35, pp. 7-26.

Kalton, G. (1981), *Compensating for Missing Survey Data* , Ann Arbor, MI: Institute for Social Research.

Kane, E., and Macaulay, L. (1993), “Interviewer Gender and Gender Attitudes,” *Public Opinion Quarterly* , 57, pp. 1-

28.

Katz, J. (1972), *Experimenting with Human Beings* , New York: Russell Sage Foundation.

Kish, L. (1949), “A Procedure for Objective Respondent Selection Within the Household,” *Journal of the American Statistical Association* , 44, pp. 380-387.

Kish, L. (1962), “Studies of Interviewer Variance for Attitudinal Variables.” *Journal of the American Statistical Association* , 57, pp. 92-115.

Kish, L. (1965), *Survey Sampling* , New York: Wiley.

Kish, L. (1988), “Multipurpose Sample Designs,” *Survey Methodology* , 14, pp. 19-32.

Kish, L. and Frankel, M. (1974), “Inference from Complex Samples” (with discussion), *Journal of the Royal Statistical Society* , Set. B, 36, pp. 1-37.

Kish, L., and Hess, I. (1959), “Some Sampling Techniques for Continuing Survey Operations,” *Proceedings of the Social Statistics Section* , *American Statistical Association* , pp. 139-143.

Kish, L., Groves, R., and Krotki, K. (1976), *Sampling Errors for Fertility Surveys* , World Fertility Survey

Occasional Paper 17, The Hague, Voorburg: International Statistical Institute.

Konnendi, E., and Noordhoek, J. (1989), *Data Quality and Telephone Surveys*, Copenhagen: Danmark's Statistik.

Krazit, T. (2001), "Like Father, Like Son," *Public Perspective*, 13, pp. 13-16.

Krosnick, J. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, pp. 213-236.

Krosnick, J. (1999), "Survey Research," *Annual Review of Psychology*, 50, pp. 537-567.

Krosnick, J. (2002), "The Causes of No-Opinion Responses to Attitude Measures in Surveys: They Are Rarely What They Appear to Be," in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse*, pp. 87-100, New York: Wiley.

Krosnick, J., and Alwin, D. (1987), "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement," *Public Opinion Quarterly*, 51, pp. 201-219.

Krosnick, J., and Berent, M. (1993), "Comparisons of Party Identification and Policy Preferences: The Impact of

Survey Question Format," *American Journal of Political Science* , 37, pp. 941-964.

Krosnick, J., and Fabrigar, L. (1997), "Designing Rating Scales for Effective Measurement in Surveys," in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality* , pp. 141-164, New York: Wiley.

Krueger, R., and Casey, M. (2000), *Focus Groups : A Practical Guidefor Applied Research* , Beverly Hills, CA: Sage Publications.

Kuusela, V., Callegaro, M., and Vehovar, V. (2008), "The Influence of Mobile Telephones on Telephone Surveys," Chapter 4 in Lepkowski, J., Tucker, C., Brick, J., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (eds.), *Advances in Telephone Survey Methodology* , pp. 87-112, New York: Wiley.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics* , 9, pp. 313-331.

Larson, R., and Richards, M. (1994), *Divergent Realities: The Emotional Lives of Mothers , Fathers , and Adolescents* , New York: Basic Books.

Lepkowski, J, Sadosky, S., and Weiss, P. (1998), "Mode, Behavior, and Data Recording Error," in Couper, M., Baker,

R., Bethlehem, J., Clark, C., Martin, J., Nicholls II, W., and O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection* , pp. 367-388, New York: Wiley.

Lepkowski, J., and Groves, R. (1986), "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design," *Journal of the American Statistical Association* , 81, pp. 930-937.

Lepkowski, J., Tucker, C., Brick, J. M., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (2008), *Advances in Telephone Survey Methodology* , New York: Wiley.

Lessler, J., Caspar, R., Penne, M., and Barker, P. (2000), "Developing ComputerAssisted Interviewing (CAI) for the National Household Survey on Drug Abuse," *Journal of Drug Issues* , 30, pp. 19-34.

Lessler, J., and Forsyth, B. (1996), "A Coding System for Appraising Questionnaires," in Schwarz, N., and Sudman, S. (eds.), *Answering Questions* , pp. 259-292, San Francisco: Jossey-Bass.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys* , New York: Wiley.

Levy, P., and Lemeshow, S., (2008), *Sampling of Populations* : *Methods and Applications* , 4th Edition, New York: Wiley.

Lievesley, D. (1988), “Unit Non-Response in Interview Surveys,” London: Social and Community Planning Research, unpublished working paper.

Likert, R. (1932), “A Technique for Measurement of Attitudes,” *Archives of Psychology* , 140, pp. 5-53.

Link, M., and Mokdad, A. (2005), “Alternative Modes for Health Surveillance Surveys: An Experiment with Web, Mail, and Telephone,” *Epidemiology* , 16, pp. 701-704.

Link, M., and Mokdad, A.H. (2006), “Can Web and Mail Survey Modes Improve Participation in an RDD-Based National Health Surveillance?” *Journal of Official Statistics* , 22, pp. 293-312.

Linton, M. (1982), “Transformations of Memory in Everyday Life,” in Neisser, U. (ed.), *Memory Observed* , pp. 77-91, San Francisco: Freeman.

Little, R., and Rubin, D. (2002), *Statistical Analysis with Missing Data* , 2nd Edition, New York: Wiley.

Little, R. J. (1993). “Statistical Analysis of Masked Data,” *Journal of Official Statistics* , 9, pp. 407-426.

Lord, F, and Novick, M. (1968), *Statistical Theories of Mental Test Scores* , Reading, MA: Addison-Wesley.

Lozar-Manfreda, K., Bosnjak, M., Haas, I., and Vehovar, V. (2008), "Web Surveys Versus Other Survey Modes: A Meta-Analysis Comparing Response Rates," *International Journal of Market Research*, 50, pp. 79-104.

Lyberg L, Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (eds.) (1997), *Survey Measurement and Process Quality*, New York: Wiley.

Mangione, T., Fowler, F, and Louis, T. (1992), "Question Characteristics and Interviewer Effects," *Journal of Official Statistics*, 8, pp. 293-307.

Mangione, T., Hingson, R., and Barrett, J. (1982), "Collecting Sensitive Data: A Comparison of Three Survey Strategies," *Sociological Methods and Research*, 10, pp. 337-346.

Martin, E. (1999), "Who Knows Who Lives Here? Within-Household Disagreeements as a Source of Survey Coverage Error," *Public Opinion Quarterly*, 63, pp. 220-236.

Martin, E. (2006), "Privacy Concerns and the Census Long From: Some Evidence from Census 2000," #2006-10 in Research Report Series, U.S. Census Bureau (http://www.census.gov/srd/papers/pdf/rsm2006-10.pdf).

Martin, J., O'Muircheartaigh, C., and Curtice, J. (1993), "The Use of CAPI for Attitude Surveys: An Experimental

Comparison with Traditional Methods," *Journal of Official Statistics* , 9, pp. 641-662.

Martinson, B., Anderson, M., and de Vries, R. (2005), "Scientisits Behaving Badly." *Nature* , 435, June 9, pp. 737-738.

Matschinger, H., Bemert, S., and Angermeyer, M. (2005), "An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview," *Journal of Official Statistics* 21, pp. 657-674.

Maynard, D., Houtkoop-Steenstra, H., Schaeffer, N., and van der Zouwen, H. (eds.) (2002), *Standardization and Tacit Knowledge* : *Interaction and Practice in the Survey Interview* , New York: Wiley.

McCabe, S., Boyd, C., Couper, M., Crawford, S., and d'Arcy, H. (2002), "Mode Effects for Collecting Alcohol and Other Drug Use Data: Web and US Mail," *Journal of Studies on Alcohol* , 63, pp. 755-761.

McHomey, C., Kosinski, M., and Ware, J. (1994), "Comparison of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail Versus Telephone Interview: Results from a National Survey," *Medical Care* , 32, pp. 551-567.

Merkle, D. and Edelman, M. (2002), “Nonresponse in Exit Polls: A Comprehensive Analysis,” in Groves, R., Dillman, D., Eltinge J., and Little R. (eds.), *Survey Nonresponse* , pp. 243-258, New York: Wiley.

Merkle, D., Edelman, M., Dykeman, K., and Brogan, C. (1998), “An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error,” paper presented at the 1998 AAPOR conference.

Milgram, S. (1963), “Behavioral Study of Obedience,” *Journal of Abnormal and Social Psychology* , 67, pp. 371-378.

Miller, P., and Cannell, C. (1977), “Communicating Measurement Objectives in the Interview,” in Hirsch et al. (eds.), *Strategies for Communication Research* , pp. 127-152, Beverly Hills: Sage Publications.

Mokdad, A., Ford, E., Bowman, B., Dietz, W., Vinicor, F., Bales, V., and Marks, J. (2003), “Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001,” *Journal of the American Medical Association* , 289, pp. 76-79.

Mokdad, A., Serdula, M., Dietz, W., Bowman, B., Marks, J., and Koplan, J. (1999), “The Spread of the Obesity Epidemic in the United States,1991-1998,” *Journal of the American Medical Association* , 282, pp. 1519-1522.

Moore, J., Pascale, J., Doyle, P., Chan, A., and Griffiths, J. (2004), "Using Field Experiments to Improve Instrument Design," in Presser, S. et al. (eds.) *Questionnaire Development Evaluation and Testing Methods*, pp. 189-207, New York: Wiley.

Moore, J., Stinson, L., and Welniak, E. (1997), "Income Measurement Error in Surveys: A Review," in Sirken, M., Herrmann, D., Schechter, S., Schwarz, N., Tanur, J., and Tourangeau, R. (eds.), *Cognition and Survey Research*, pp. 155-174, New York: Wiley.

Morton-Williams, J. (1993), *Interviewer Approaches*, Aldershot, U.K.: Dartmouth.

Mulry, M. (2007), "Summary of Accuracy and Coverage Evaluation for the U. S. Census 2000," *Journal of Official Statistics*, 23, pp. 345-370.

National Bioethics Advisory Commission (2001), *Ethical and Policy Issues in Research Involving Human Participants*, Vol. 1: Report and Recommendations, Bethesda, MD: National Bioethics Advisory Commission.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979), *Belmont Report: Ethical Principles and Guidelines for the Protection of*

Human Subjects of Research , Washington, DC: U. S. Government Printing Office.

National Endowment for the Arts (1998), 1997 *Survey of Public Participation in the Arts* , Research Division Report 39, Washington, DC: National Endowment for the Arts.

National Research Council (1979), *Privacy and Confidentiality as Factors in Survey Response* , Washington, DC: National Academy Press.

National Research Council (1993), *Private Lives and Public Policies* : *Confidentiality and Accessibility of Government Statistics* , Washington, DC: National Academy Press.

National Research Council (2003). *Protecting Participants and Facilitating Social and Behavioral Sciences Research* , Washington, DC: National Academy Press.

National Research Council (2006), *Expanding Access to Research Data* : *Reconciling Risks and Opportunities* , Washington, DC: National Academy Press.

Nealon, J. (1983), "The Effects of Male vs. Female Telephone Interviewers," *Proceedings of the Survey Research Methods Section of the American Statistical Association* , pp. 139-141.

Neter, J., and Waksberg, J. (1964), "A Study of Response Errors in Expenditures Data from Household Interviews," *Journal of the American Statistical Association*, 59, pp. 17-55.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, pp. 558-625.

Nicholls II, W., Baker, R., and Martin, J. (1997), "The Effect of New Data Collection Technologies on Survey Data Quality," in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C, Schwarz, N., and Trewin, D. (eds.), *Survey Measurement and Process Quality*, pp. 221-248, New York: Wiley.

O'Muircheartaigh, C. (1991), "Simple Response Variance: Estimation and Determinants," in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys*, pp. 551-574, New York: Wiley.

O'Neill, G., and Sincavage, J. (2004), "Response Analysis Survey: A Qualitative Look at Response and Nonresponse in the American Time Use Survey," retrieved December 14, 2006, from www.bls.gov/ore/pdf/st040140.pdf, Washington, DC: Bureau of Labor Statistics.

O'Toole, B., Battistutta, D., Long, A., and Crouch, K. (1986), "A Comparison of Costs and Data Quality of Three Health Survey Methods: Mail, Telephone and Personal Home Interview," *American Journal of Epidemiology* , 124, pp. 317-328.

Oksenberg, L., Cannell, C., and Kalton, G. (1991), "New Strategies of Pretesting Survey Questions," *Journal of Official Statistics* , 7, pp. 349-366.

Oksenberg, L., Coleman, L., and Cannell, C. (1986), "Interviewers' Voices and Refusal Rates in Telephone Surveys," *Public Opinion Quarterly* , 50, pp. 97-111.

Olson, K. (2006), "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias, *Public Opinion Quarterly* , 70, pp. 737-758.

Pastore, A. and Maguire, K. (eds.) (2008), *Sourcebook of Criminal Justice Statistics* [Online]. Available at http://www.albany.edu/sourcebook/.

Payne, S. (1951), *The Art of Asking Questions* , Princeton, NJ: Princeton University Press.

Pearson, R., Ross, M., and Dawes, R. (1992), "Personal Recall and the Limits of Retrospective Questions in Surveys," in Tanur, J. (ed.), *Questions About Questions:*

Inquiries into the Cognitive Basis of Surveys , pp. 65-94, New York: Russel Sage.

Pillemer, D. (1984), “Flashbulb Memories of the Assassination Attempt on President Reagan,” *Cognition* , 16, pp. 63-80.

Presser, S. (1990), “Measurement Issues in the Study of Social Change,” *Social Forces* , 68, pp. 856-868.

Presser, S. (1994), “Informed Consent and Confidentiality in Survey Research,” *Public Opinion Quarterly* , 58, pp. 446-459.

Presser, S. and Blair, J. (1994), “Survey Pretesting: Do Different Methods Produce Different Results?” in Marsden, P. (ed.), *Sociology Methodology* , 24, pp. 73-104, Washington DC: American Sociological Association.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E. (eds.) (2004), *Methods for Testing and Evaluating Survey Questionnaires* , New York: Wiley.

Rand, M. and Rennison, C. (2002), “True Crime Stories? Accounting for Differences in our National Crime Indicators,” *Chance* , 15, pp. 47-51.

Rasinski, K. (1989), “The Effect of Question Wording on Support for Government Spending,” *Public Opinion Quarterly* , 53, pp. 388-394.

Rasinski, K., Mingay, D., and Bradburn, N. (1994), “Do Respondents Really ‘Mark All That Apply’ on Self-Administered Questions?” *Public Opinion Quarterly* , 58, pp. 400-408.

Redline, C., and Dillman, D. (2002), “The Influence of Alternative Visual Designs on Respondents’ Performance with Branching Instructions in SelfAdministered Questionnaires,” in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 179-195, New York: Wiley.

Richman, W., Kiesler, S., Weisband, S., and Drasgow, F. (1999), “A MetaAnalytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaires, and Interviews,” *Journal of Applied Psychology* , 84, pp. 754-775.

Rizzo, L., Brick, J., and Park, I. (2004), “A Minimally Intrusive Method for Sampling Persons in Random-Digit Dial Surveys,” *Public Opinion Quarterly* , 68, pp. 267-274.

Robinson, D., and Rohde, S. (1946), “Two Experiments in an Anti-Semitism Poll,” *Journal of Abnormal and Social Psychology* , 41, pp. 136-144.

Robinson, J., Ahmed, B., das Gupta, P., and Woodrow, K. (1993), "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association*, 88, pp. 1061-1071.

Robinson, J., Neustadtl, A., and Kestnbaum, M. (2002), "Why Public Opinion Polls Are Inherently Biased: Public Opinion Differences Among Internet Users and Non-Users," paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL, May.

Rosenthal, R., and Rosnow, R. (1975), *The Volunteer Subject*, New York: Wiley.

Rothgeb, J., Willis, G., and Forsyth, B. (2001), "Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?" in *Proceedings of the Survey Research Methods Section*, http://www.amstat.org/sections/SRMS/Proceedings/.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. (1993), "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics*, 9, pp. 461-468.

Rubin, D., and Baddeley, A. (1989), “Telescoping is Not Time Compression: A Model of the Dating of Autobiographical Events,” *Memory and Cognition* , 17, pp. 653-661.

Rubin, D., and Kozin, M. (1984), “Vivid Memories,” *Cognition* , 16, pp. 81-95.

Rubin, D., and Wetzel, A. (1996), “One Hundred Years of Forgetting: A Quantitative Description of Retention,” *Psychological Review* , 103, pp. 734-760.

Saris, W., and Andrews, F. (1991), “Evaluation of Measurement Instruments Using a Structural Modeling Approach,” in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 575-597, New York: Wiley.

Saris, W., and Gallhofer, I. (2007), *Design* , *Evaluation, and Analysis of Questionnaires for Survey Research.* , New York: Wiley.

Särndal, C., and Lundström, S. (2005), *Estimation in Surveys with Nonresponse* , New York: Wiley.

Schaefer, C., Schraepler, J-P., Mueller, K., and Wagner, G. (2005), “Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methods,” in *Proceedings of the Survey Research Methods Section* , *American*

Statistical Association , pp. 4318-4325, Alexandria, VA, American Statistical Association.

Schaefer, D., and Dillman, D. (1998), "Development of a Standard E-Mail Methodology: Results of an Experiment," *Public Opinion Quarterly* , 62, pp. 378-397.

Schaeffer, N. (1991), "Conversation with a Purpose or Conversation? Interaction in the Standardized Interview," in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 367-393, New York: Wiley.

Schaeffer, N., and Bradburn, N. (1989), "Respondent Behavior in Magnitude Estimation," *Journal of the American Statistical Association* , 84, pp. 402-413.

Schober, M., and Conrad, F. (1997), "Does Conversational Interviewing Reduce Survey Measurement Error," *Public Opinion Quarterly* , 61, pp. 576-602.

Schober, S., Caces, M., Pergamit, M., and Branden, L. (1992), "Effects of Mode of Administration on Reporting of Drug Use in the National Longitudinal Survey," in Turner, C., Lessler, J., and Gfroerer, J. (eds.), *Survey Measurement of Drug Use : Methodological Studies* , pp. 267-276, Rockville, MD: National Institute on Drug Abuse.

Schreiner, I., Pennie, K., and Newbrough, J. (1988), "Interviewer Falsification in Census Bureau Surveys," in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 491-496, Washington, DC: American Statistical Association.

Schuman, H. (1997), "Polls, Surveys, and the English Language," *The Public Perspective*, April/May, pp. 6-7.

Schuman, H., and Converse, J. (1971), "The Effects of Black and White Interviewers on Black Responses in 1968," *Public Opinion Quarterly*, 35, pp. 44-68.

Schuman, H., and Hatchett, S. (1976), "White Respondents and Race-of Interviewer Effects," *Public Opinion Quarterly*, 39, pp. 523-528.

Schuman, H., and Presser, S. (1981), *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context*, New York: Academic Press.

Schwarz, N., and Sudman, S. (eds.) (1992), *Context Effects in Social and Psychological Research*, New York: Springer.

Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985), "Response Categories: Effects on Behavioral Reports and Comparative Judgments," *Public Opinion Quarterly*, 49, pp. 388-395.

Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, F. (1991), “Rating Scales: Numeric Values May Change the Meaning of Scale Labels,” *Public Opinion Quarterly* , 55, pp. 618-630.

Schwarz, N., Strack, F., Hippler, H., and Bishop, G. (1991), “The Impact of Administration Mode on Response Effects in Survey Measurement,” *Applied Cognitive Psychology* , 5, pp. 193-212.

Schwarz, N., Strack, F., and Mai, H. (1991), “Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis,” *Public Opinion Quarterly* , 55, pp. 3-23.

Sheppard, J. (2001), 2001 *Respondent Cooperation and Industry Image Study* : *Privacy and Survey Research* , Council of Marketing and Opinion Research, Cincinnati, OH: CMOR.

Shih, T.-H., and Fan, X. (2008), “Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis,” *Field Methods* , 20, pp. 249-271.

Short, J., Williams, E., and Christie, B. (1976), *The Social Psychology of Telecommunications* , New York: Wiley.

Singer, E. (1978), “Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys,” *American Sociological Review* , 43, pp. 144-162.

Singer, E. (1984), “Public Reactions to Some Ethical Issues of Social Research: Attitudes and Behavior,” *Journal of Consumer Research* , 11, pp. 501-509.

Singer, E. (2002), “The Use of Incentives to Reduce Nonresponse in Household Surveys,” in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 163-177, New York: Wiley.

Singer, E. (2003), “Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits,” *Journal of Official Statistics* , 19, pp. 273-285.

Singer, E., and Bossarte, R. (2006), “Incentives for Survey Participation: When Are They Coercive?” *American Journal of Preventive Medicine* , 31, pp. 411-418.

Singer, E., and Couper, M. (2008), “Do Incentives Exert Undue Influence on Survey Participation? Experimental Evidence,” *Journal of Empirical Research on Human Research Ethics* , 3, pp. 49-56.

Singer E., and Couper, M., “Ethical Considerations in Internet Surveys,” forth coming in L. Kaczmirek, M. Das, and P. Ester, eds., *Social Research and the Internet.*

Singer, E., and Frankel, M. (1982), “Informed Consent in Telephone Interviews,” *American Sociological Review* , 47, pp. 116-126.

Singer, E., Mathiowetz, N., and Couper, M. (1993), “The Role of Privacy and Confidentiality as Factors in Response to the 1990 Census,” *Public Opinion Quarterly* , 57, pp. 465-482.

Singer, E., and Presser, S. (2007), “Privacy, Confidentiality, and Respondent Burden as Factors in Telephone Survey Nonresponse,” in Lepkowski, J., Tucker, C., Brick, J., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R. (eds.), *Advances in Telephone Survey Methodology* , New York: Wiley.

Singer, E., Groves, R., and Coming, A. (1999), “Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation,” *Public Opinion Quarterly* , 63, pp. 251-260.

Singer, E., Hippler, H-J., and Schwarz, N. (1992), “Confidentiality Assurances in Surveys: Reassurance or Threat,” *International Journal of Public Opinion Research* , 4, pp. 256-68.

Singer, E., Van Hoewyk, J., and Maher, M. (2000), “Experiments with Incentives in Telephone Surveys,” *Public Opinion Quarterly* , 64, pp. 171-188.

Singer, E., Van Hoewyk, J., and Neugebauer, R. (2003), “Attitudes and Behavior: The Impact of Privacy and

Confidentiality Concerns on Participation in the 2000 Census," *Public Opinion Quarterly* , 65, pp. 368-384.

Singer, E., Von Thurn, D., and Miller, R. (1995), "Confidentiality Assurances and Survey Response: A Review of the Experimental Literature," *Public Opinion Quarterly* , 59, pp. 266-277.

Singleton, R., and Straits, B. (2005). *Approaches to Social Research* , 4th edition, New York: Oxford University Press.

Sirken, M. (1970), "Household Surveys with Multiplicity," *Journal of the American Statistical Association* , 65, pp. 257-266.

Smith, A. (1991), "Cognitive Processes in Long-Term Dietary Recall," *Vital and Health Statistics* , Series 6, No. 4 (DHHS Publication No. PHS 92-1079), Washington, DC: U.S. Government Printing Office.

Smith, M. (1979), "Some Perspectives on Ethical/Political Issues in Social Science Research," in Wax, M., and Cassell, J. (eds.), *Federal Regulations : Ethical Issues and Social Research* , pp. 11-22, Boulder, CO: Westview Press.

Smith, P., MacQuarrie, C., Herbert, R., Cairns, D., and Begley, L. (2004), "Preventing Data Fabrication in Telephone

Survey Research," *Journal of Research Administration* , 35, pp. 13-21.

Smith, T. (1983), "The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey," *Public Opinion Quarterly* , 47, pp. 386-404.

Smith, T. (1984), "Recalling Attitudes: An Analysis of Retrospective Questions on the 1982 General Social Survey," *Public Opinion Quarterly* , 48, pp. 639-649.

Smith, T. (1987), "That Which We Call Welfare By Any Other Name Would Smell Sweeter: An Analysis of the Impact of Question Wording on Response Patterns," *Public Opinion Quarterly* , 51, pp. 75-83.

Sperry, S., Edwards, B., Dulaney, R., and Potter, D. (1998), "Evaluating Interviewer Use of CAPI Navigation Features," in Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls II, W., and O'Reilly, J. (eds.), *Computer-Assisted Survey Information Collection* , pp. 351-366, New York: Wiley.

Steiger, D., and Conroy, B. (2008), "IVR: Interactive Voice Response," in de Leeuw, E. D., Hox, J. J., and Dillman, D. A (eds.), *International Handbook of Survey Methodology.* New York: Lawrence Erlbaum, pp. 285-298.

Stewart, A., Ware, J., Sherbourne, C., and Wells, K. (1992), “Psychological Distress/Well-Being and Cognitive Functioning Measures,” in Stewart, A., and Ware, J. (eds.), *Measuring Functioning and Well-Being* : *The Medical Outcomes Study Approach* , pp. 102-142, Durham, NC: Duke University Press.

Suchman, L., and Jordan, B. (1990), “Interactional Troubles in Face-to-Face Survey Interviews,” *Journal of the American Statistical Association* , 85, pp. 232-241.

Sudman, S., and Bradburn, N. (1982), *Asking Questions* : *A Practical Guide to Questionnaire Design* , San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N., and Schwarz, N. (1996), *Thinking About Answers* : *The Application of Cognitive Processes to Survey Methodology* , San Francisco: Jossey-Bass.

Swazey, J., Anderson, M., and Lewis, K. (1993), “Ethical Problems in Academic Research,” *American Scientist* , 81, pp. 542-553.

Sykes, W., and Collins, M. (1988), “Effects of Mode of Interview: Experiments in the UK,” in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 301-320, New York: Wiley.

Taylor, B., and Rand, M. (1995), "*The National Crime Victimization Survey Redesign : New Understandings of Victimization Dynamics and Measurement* ," paper prepared for presentation at the 1995 American Statistical Association Annual Meeting, August 13-17, 1995 in Orlando, Florida (http://www.ojp.usdoj.gov/bjs/ncvsrd96.txt).

Tarnai, J. and Dillman, D. (1992), "Questionnaire Context as a Source of Response Differences in Mail vs. Telephone Surveys," in Schwarz, N., and Sudman, S. (eds.), *Context Effects in Social and Psychological Research* , pp. 115-129, New York: Springer-Verlag.

Tarnai, J., and Moore, D. (2004), "Methods for Testing and Evaluating CAI Questionnaires," in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Thornberry, O., and Massey, J. (1988), "Trends in United States Telephone Coverage Across Time and Subgroups," in Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J. (eds.), *Telephone Survey Methodology* , pp. 25-50, New York: Wiley.

Thurstone, L., and Chave, E. (1929), *The Measurement of Attitude* , Chicago: University of Chicago.

Titus, S., Wells, J., and Rhoades, L. (2008), "Repairing Research Integrity," *Nature* , 453, June 19, pp. 980-982.

Tourangeau, R. (1984), "Cognitive Science and Survey Methods," in Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (eds.), *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines* , pp. 73-100, Washington, DC: National Academy Press.

Tourangeau, R. (2004), "Design Considerations for Questionnaire Testing and Evaluation," in Presser, S. et al. (eds.), *Questionnaire Development Evaluation and Testing Methods* , New York: Wiley.

Tourangeau, R., Rasinski, K., Jobe, J., Smith, T., and Pratt, W. (1997), "Sources of Error in a Survey of Sexual Behavior," *Journal of Official Statistics* , 13, pp. 341-365.

Tourangeau, R., Rips, L., and Rasinski, K. (2000), *The Psychology of Survey Response* , Cambridge: Cambridge University Press.

Tourangeau, R., Shapiro, G., Kearney, A., and Ernst, L. (1997), "Who Lives Here? Survey Undercoverage and Household Roster Questions," *Journal of Official Statistics* , 13, pp. 1-18.

Tourangeau, R., and Smith, T (1996), "Asking Sensitive Questions: The Impact of Data Collection, Question Format,

and Question Context,” *Public Opinion Quarterly* , 60, pp. 275-304.

Tourangeau, R., Steiger, D., and Wilson, D. (2002), “Self-Administered Questions by Telephone: Evaluating Interactive Voice Response,” *Public Opinion Quarterly* , 66, pp. 265-278.

Traugott, M., Groves, R., and Lepkowski, J. (1987), “Using Dual Frame Designs to Reduce Nonresponse in Telephone Surveys,” *Public Opinion Quarterly* , 51, pp. 522-539.

Trice, A. (1987), “Informed Consent: Biasing of Sensitive Self-Report Data by Both Consent and Information,” *Journal of Social Behavior and Personality* , 2, pp. 369-374.

Tucker, C., Lepkowski, J., and Piekarski, L. (2002), “List-Assisted Telephone Sampling Design Efficiency,” *Public Opinion Quarterly* , 66, pp. 321-338.

Turner, C., Forsyth, B., O'Reilly, J., Cooley, P., Smith, T., Rogers, S., and Miller, H. (1998), “Automated Self-interviewing and the Survey Measurement of Sensitive Behaviors,” in Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls II, W., and O'Reilly, J. (eds.), *Computer Assisted Survey Information Collection* , pp. 455-473, New York: Wiley.

Turner, C., Lessler, J., and Devore, J. (1992), "Effects of Mode of Administration and Wording on Reporting of Drug Use," in Turner, C., Lessler, J., and Gfroerer, J. (eds.), *Survey Measurement of Drug Use : Methodological Studies* , pp. 177-220, Rockville, MD: National Institute on Drug Abuse.

Turner, C., Lessler, J., George, B., Hubbard, M., and Witt, M. (1992), "Effects of Mode of Administration and Wording on Data Quality," in Turner, C., Lessler, J., and Gfroerer, J. (eds.), *Survey Measurement of Drug Use: Methodological Studies* , pp. 221-244, Rockville, MD: National Institute on Drug Abuse.

Turner, C., Lessler, J., and Gfroerer, J. (1992), *Survey Measurement of Drug Use : Methodological Studies* , Washington, DC: National Institute on Drug Abuse.

Tuskegee Syphilis Study Ad Hoc Advisory Panel, (1973), *Final Report* , Washington, DC: U. S. Department of Health, Education, and Welfare.

U. S. Bureau of Justice Statistics (1994), *National Crime Victimization Survey (NCVS) Redesign : Questions and Answers* , NCJ 151171, Washington, DC: Bureau of Justice Statistics.

U. S. Bureau of the Census (1993), "Memorandum for Thomas C. Walsh from John H. Thompson, Subject: 1990 Decennial Census—Long Form (Sample Write-in) Keying

Assurance Evaluations," Washington, DC: U.S. Bureau of the Census.

U.S. Bureau of Labor Statistics (2003), *BLS Handbook of Methods* , http://www.bls.gov/opub/hom/home.htm.

U. S. Bureau of Labor Statistics, *Monthly Labor Review* , http://www.bls.gov/opub/mlr/mlrhome.htm.

U. S. Census Bureau (2003), "Noninterview Rates for Selected Major Demographic Household Surveys," memorandum from C. Bowie, August 25, Xerox.

U. S. Census Bureau (2006) *Voting and Registration in the Election of* 2004, Current Population Reports, pp. 20-556, Washington, DC: U.S. Census Bureau.

U. S. Department of Education. National Center for Education Statistics (2003), *NCES Handbook of Survey Methods* , NCES 2003-2603, by Lori Thurgood, Elizabeth Walter, George Carter, Susan Henn, Gary Huang, Daniel Nooter, Wray Smith, R. William Cash, and Sameena Salvucci. Project Officers, Marilyn Seastrom, Tai Phan, and Michael Cohen. Washington, DC.

U. S. Dept. of Health, Education, and Welfare (1974), "Protection of Human Subjects," *Federal Register* , 39 (105), May 30, Pt. II, pp. 18914-18920.

van Campen, C., Sixma, H., Kerssens, J., and Peters L. (1998), "Comparisons of the Costs and Quality of Patient Data Collection by Mail Versus Telephone Versus In-Person Interviews," *European Journal of Public Health* , 8, pp. 66-70.

van der Zouwen, J., Dijkstra, W., and Smit, J. (1991), "Studying RespondentInterviewer Interaction: The Relationship between Interviewing Style, Interviewer Behavior, and Response Behavior," in Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.), *Measurement Errors in Surveys* , pp. 419--438, New York: Wiley.

van Leeuwen, R., and de Leeuw, E. (1999), "I Am Not Selling Anything: Experiments in Telephone Introductions," paper presented at the International Conference on Survey Nonresponse, Portland, OR.

Vehovar, v., Batagelj, Z., Lozar Manfreda, K., and Zaletel, M. (2002), "Nonresponse in Web Surveys," in Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.), *Survey Nonresponse* , pp. 229-242, New York: Wiley.

Vinovskis, M. (1998), *Overseeing the Nation's Report Card : The Creation and Evolution of the National Assessment Governing Board* (NAGB), Washington, DC: U. S. Government Printing Office.

Wagenaar, W. (1986), “My Memory: A Study of Autobiographical Memory Over Six Years,” *Cognitive Psychology* , 18, pp. 225-252.

Walker, A., and Restuccia, J. (1984), “Obtaining Information on Patient Satisfaction with Hospital Care: Mail Versus Telephone,” *Health Services Research* , 19, pp. 291-306.

Warner, J., Berman, J., Weyant, J., and Ciarlo, J. (1983), “Assessing Mental Health Program Effectiveness: A Comparison of Three Client Follow-up Methods,” *Evaluation Review* , 7, pp. 635-658.

Warner, S. (1965), “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” *Journal of the American Statistical Association* , 60, pp. 63-69.

Weeks, M., Kulka, R., Lessler, J., and Whitmore, R. (1983), “Personal Versus Telephone Surveys for Collecting Household Health Data at the Local Level,” *American Journal of Public Health* , 73, pp. 1389-1394.

Weisberg, H., (2005), *The Total Survey Error Approach* : *A Guide to the New Science ofSurvey Research* , Chicago: The University of Chicago Press.

Weiss, C. (1968), “Validity of Welfare Mothers’ Interview Responses,” *Public Opinion Quarterly* , 32, pp.

622-633.

Wells, G. (1993), “What Do We Know About Eyewitness Identification?” *American Psychologist* , 48, pp. 553-571.

Willis, G., DeMaio, T., and Harris-Kojetin, B. (1999), “Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques,” in Sirken, M., et al. (eds.), *Cognition in Survey Research* , pp. 133-154, New York: Wiley.

Willis, G., Schechter, S., and Whitaker, K. (2000), “A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What Do They Tell Us?” in *Proceedings of the Section on Survey Research Methods* , *American Statistical Association* , pp. 28-37. Alexandria, VA: American Statistical Association.

Willis, G. (2005), *Cognitive Interviewing* : *A Tool for Improving Questionnaire Design* , Thousand Oaks, CA: Sage.

Wilson, T., and Hodges, S. (1992), “Attitudes as Temporary Constructions,” in Martin, L., and Tesser, A. (eds.), *The Construction of Social Judgments* , pp. 37-66，New York: Springer-Verlag.

Witte, J., Amoroso, L., and Howard, P. (2000), “Method and Representation in Internet-Based Survey Tools: Mobility,

Community, and Cultural Identity in Survey2000." *Social Science Computer Review* , 18, pp. 179-195.

Yu, J., and Cooper, H. (1983), "A Quantitative Review of Research Design Effects on Response Rates to Questionnaires," *Journal of Marketing Research* , 20, pp. 36-44.

Zayatz, L. (2007), "Disclosure Avoidance Practices and Research at the U. S. Census Bureau: An Update," presented at Workshop on Ensuring Access and Confidentiality Protection for Highly Sensitive Data, Institute for Social Research, University of Michigan, October 3.